

DEFENSE EQUAL OPPORTUNITY MANAGEMENT INSTITUTE

DIRECTORATE OF RESEARCH

Comparing Two Versions of the MEOCS Using Differential Item Functioning

by

Stephen A. Truhon, Ph.D
Winston-Salem State University

Summer 2002

20030917 063

DEOMI Research Report 02-07



REPORT DOCUMENTATION PAGE					Form Approved OMB No. 0704-0188	
The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.						
1. REPORT DATE (DD-MM-YYYY) 20-05-2003		2. REPORT TYPE Research		3. DATES COVERED (From - To) June - August 2003		
4. TITLE AND SUBTITLE Comparing Two Versions of the MEOCS Using Differential Item Functioning				5a. CONTRACT NUMBER N00014-97-1-1055		
				5b. GRANT NUMBER N/A		
				5c. PROGRAM ELEMENT NUMBER N/A		
				5d. PROJECT NUMBER N/A		
6. AUTHOR(S) Stephen A. Truhon				5e. TASK NUMBER N/A		
				5f. WORK UNIT NUMBER N/A		
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Directorate of Research Defense Equal Opportunity Management Institute 740 O'Malley Road MS9121 Patrick Air Force Base, FL 32925-3399				8. PERFORMING ORGANIZATION REPORT NUMBER DEOMI RESEARCH REPORT 02-07		
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Office of Naval Research 800 N. Quincy Street Arlington, VA 22032				10. SPONSOR/MONITOR'S ACRONYM(S) ONR		
				11. SPONSOR/MONITOR'S REPORT NUMBER(S) N/A		
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution unlimited.						
13. SUPPLEMENTARY NOTES						
14. ABSTRACT As part of a project to update the Military Equal Opportunity Climate Survey (MEOCS), items from eight of its scales (Sexual Harassment and Discrimination, Differential Command Behavior toward Minorities and Women, Positive Equal Opportunity (EO) Behavior, Racist/Sexist Behavior, Reverse Discrimination (Behavior), Discrimination against Minorities and Women, Reverse Discrimination (Attitude), and Attitudes toward Racial/Gender Separatism) have been rewritten to make them more neutral (e.g., replacing terms such as "men" and "women" with "gender"). A comparison was made of the psychometric characteristics of the original and revised versions of these items using differential item functioning (DIF) from item response theory (IRT). DIF was found for the majority of the 40 items examined, although in many cases the DIF indicated improvements in the revised items. Implications for these scales and for the use of IRT with the MEOCS are discussed.						
15. SUBJECT TERMS Equal Opportunity, Climate Survey, Military Climate, Discrimination, Differential Item Functioning, Item Response theory						
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT		18. NUMBER OF PAGES	
a. REPORT	b. ABSTRACT	c. THIS PAGE	UU		19a. NAME OF RESPONSIBLE PERSON Jerry C. Scarpate	
U	U	U			19b. TELEPHONE NUMBER (Include area code) (321) 494-2676	

Comparing Two Versions of the MEOCS Using Differential Item Functioning

Stephen A. Truhon, Ph.D.
Associate Professor of Psychology
Department of Social Sciences
Winston-Salem State University

Abstract

As part of a project to update the Military Equal Opportunity Climate Survey (MEOCS), items from eight of its scales (Sexual Harassment and Discrimination, Differential Command Behavior toward Minorities and Women, Positive Equal Opportunity (EO) Behavior, Racist/Sexist Behavior, Reverse Discrimination (Behavior), Discrimination against Minorities and Women, Reverse Discrimination (Attitude), and Attitudes toward Racial/Gender Separatism) have been rewritten to make them more neutral (e.g., replacing terms such as “men” and “women” with “gender”). A comparison was made of the psychometric characteristics of the original and revised versions of these items using differential item functioning (DIF) from item response theory (IRT). DIF was found for the majority of the 40 items examined, although in many cases the DIF indicated improvements in the revised items. Implications for these scales and for the use of IRT with the MEOCS are discussed.

Summer 2002

Opinions expressed in this report are those of the author and should not be construed to represent the official position of DEOMI, the military Services, or the Department of Defense.

Comparing Two Versions of the MEOCS Using Differential Item Functioning

Stephen A. Truhon, Ph.D.
Winston-Salem State University

Executive Summary¹

Item response theory (IRT) arose, in part, to meet some of the failings of classical test theory (CTT). CTT is concerned with the measurement of a trait or ability by a test. The score that a person obtains on that test is a linear combination of the person's true trait or ability and a certain amount of error. Even with that score, the person's ability is defined in terms of the particular test given. If the person takes a difficult ability test, he or she may appear to have less of that ability than if he or she had taken an easy test. Frequently a person's score is compared to others' scores. A person of moderate ability may appear deficient if the comparison group consists of those with high ability; likewise, he or she may appear superior if the comparison group consists of those with low ability. A further problem occurs when a person's score is compared with someone who has taken a different form of the test. A comparison can be made if the tests are parallel; but parallelism is difficult to obtain.

IRT assumes that the relationship between a person's score on a test and his or her true ability is nonlinear, more specifically expressed by an ogive or S-shaped function. IRT defines the person's trait in terms of the items used. There is a constant process of estimating abilities by including information about the items and estimating the difficulty of items by including information about the persons' abilities are chosen. Tests can be adapted to the individual so items whose difficulty is close to the person's ability. Thus, the person does not need to be compared to others and individuals can take different forms of a test without worrying about parallelism.

There are several models for use in IRT. One model of interest in using the MEOCS is the two-parameter logistic model. Using the ogive function mentioned earlier, two parameters are calculated: the discrimination parameter (a), which represents the slope of the ogive, and the difficulty parameter (b), which represents the level of the ability required to have a 50% chance of getting an item correct.

The two-parameter logistic model can be extended for the MEOCS. Early IRT models made use of dichotomous responses (i.e., right-wrong, true-false). Many attitude scales like the MEOCS possess ordered polytomous responses (where there are more than two responses which range from "Strongly Disagree" to "Strongly Agree"). In such cases, a graded-response model (Samejima, 1969, 1997) is often used.

The nature of the difficulty parameter (b) changes when using tests like the MEOCS. First, difficulty is understandable when discussing ability tests. However, in attitude or personality tests, b is better defined as a between-category threshold (e.g., between "Strongly

¹ Because of the difficulty of this material for someone unfamiliar with item response theory, a synopsis of the item response theory and its meaning for the current study are provided here. This section can be read independently of the material in the remainder of the report. I thank Dr. Robert McIntyre for the suggestion.

Disagree" and "Disagree"). Second, there are multiple b 's, one less than the number of categories (i.e., in a five-point scale, there will be four b 's). For each of these b 's, there is a separate ogive called a boundary response function.

A further advantage of IRT is how it deals with item bias or, better named, differential item functioning (correspondingly, test bias or differential test functioning). In CTT, bias is said to occur when the means for two groups differ on an item (test) and when their correlations to an external variable differ (such as when the predictive validity of a test differs for two groups). While the term "bias" is often used, differential item (test) functioning is said to occur when the above difference is irrelevant to the testing (i.e., it is an artifact of the measuring process itself), rather than relevant to the testing (i.e., it reflects a real difference between the groups on that trait or ability). In addition, while item and test bias have negative connotations, differential item functioning and differential test functioning are terms that are more neutral. These terms could be viewed positively, in certain circumstances. For example, if a revised version of a test had better psychometric qualities (a higher a and b 's closer to zero) than the original, a comparison of the two tests might indicate significant differential item functioning (see Camilli & Shepard, 1994 for more detail).

The process of performing differential item functioning analysis usually involves three steps. In the first step, an IRT model is employed. In this study, the two-parameter logistic graded-response model was used through the MULTILOG program (Thissen, 1991). In the second step, the IRT analyses for the two groups need to be equated. Each IRT analysis estimates the trait/ability, discrimination, and between-category thresholds separately. To compare two groups there must be a common scale or metric. The EQUATE program (Baker, 1992) was used for this step. In the third step, the differential item functioning can be done. While it would have been preferable to use a parametric program such as DFIT (Raju, van der Linden, & Fleer, 1995), I was unable to obtain it. Therefore I used the polytomous adaptation of the SIBTEST called POLY-SIBTEST (Chang, Mazzeo, & Roussos, 1996), a nonparametric program.

As part of the process of updating the MEOCS, 40 items (five items each from the Sexual Harassment and Discrimination, Differential Command Behavior toward Minorities and Women, Positive Equal Opportunity (EO) Behavior, Racist/Sexist Behavior, Reverse Discrimination (Behavior), Discrimination against Minorities and Women, Reverse Discrimination (Attitude), and Attitudes toward Racial/Gender Separatism scales) were rewritten making their content more neutral (e.g., replacing terms "men" and "women" with "gender"). Complete responses from 1,040 participants were obtained. Their responses were compared to a random sample of 1,040 participants who had completed the corresponding items on the Standard MEOCS.

A comparison was made between the mean ratings from the two groups. There were 26 significant differences between the ratings. Nineteen of the revised items were rated in such a way to indicate a more positive equal opportunity climate than did the original versions.

All five of the items from the Sexual Harassment and Discrimination scale showed significant differential item functioning. Three items showed worse psychometric qualities in the revised versions; two items showed better psychometric qualities.

Three items from the Differential Command Behavior toward Minorities and Women scale showed significant differential item functioning. Two of these items displayed non-uniform differential item functioning² with better psychometric properties in the revised versions. The other significant item showed better psychometric properties in the revised version.

Two items from the Positive EO Behavior scale showed significant differential item functioning. These items displayed non-uniform differential item functioning with worse psychometric properties in the revised versions. The three other items all displayed non-uniform differential item functioning for all or some of the between-category threshold (*b*'s). For each of these items, the revised versions had worse psychometric properties.

All five of the items from the Racist/Sexist Behavior scale showed significant differential item functioning. Three items displayed uniform differential functioning, one with better psychometric properties for the revised version, two with worse. The other two items displayed non-uniform differential item functioning, one with better psychometric properties for the revised version, one with worse.

Three items from the Reverse Discrimination (Behavior) scale showed significant differential item functioning. These items displayed non-uniform differential item functioning for all or some of the between-category threshold (*b*'s). For each of these items, the revised versions had better psychometric properties for the revised versions. One of the other items had non-uniform differential item functioning for some of the between-category threshold (*b*'s) with better psychometric properties for the revised version.

Three items from the Discrimination against Minorities and Women scale showed significant differential item functioning. These items displayed non-uniform differential item functioning for all or some of the between-category threshold (*b*'s). For two of the items the revised versions had better psychometric properties for the revised versions, one had worse. The other two items were written identically in the original and revised versions.

All five of the items from the Reverse Discrimination (Attitude) scale showed significant differential item functioning. Three items have uniform differential item functioning, one of these with better psychometric properties for the revised version. The other two items displayed non-uniform differential item functioning for all or some of the between-category threshold (*b*'s). One of these items had better psychometric properties for the revised version.

Three items from the Attitudes toward Racial/Gender Separatism scale showed significant differential item functioning. One of these items has uniform differential item functioning with worse psychometric properties for the revised version. The other two significant items displayed non-uniform differential item functioning for some of the between-category threshold (*b*'s). One of these items had better psychometric properties for the revised version. The two nonsignificant items displayed non-uniform differential item functioning for

² Non-uniform differential item functioning occurs when the curves for two groups cross over at some point on the trait or ability scale. This situation can result in nonsignificant differential item functioning if the positive and negative differences cancel each other out (Camilli & Shepard, 1994, pp. 59-60).

all or some of the between-category threshold (b 's). One of these items had better psychometric properties for the revised version.

In general, many of the items displayed significant differential item functioning when comparing the original versions. Roughly half of these significant findings indicated that the revised version of an item had better psychometric properties. These findings should be viewed cautiously and compared with the results from a parametric differential item functioning program, such as DFIT (Raju et al., 1995).

IRT and differential item functioning provide tools that can be useful in the further development of the MEOCS. IRT has proven useful in test construction through the development of new items for a test, the rewriting of items for a test, the development of alternate forms of a test, and the shortening of scales. Differential item functioning techniques can be used to examine differences between gender and racial/ethnic groups and eliminate item and test bias. Finally, IRT is useful in computerized adaptive testing (CAT). CAT is most frequently done with ability testing, but there are instances of it being used with personality and attitude tests.

Introduction

The major research project for the Defense Equal Opportunity Management Institute (DEOMI) has been the development and testing of the Military Equal Opportunity Climate Survey (MEOCS; Landis, Dansby, & Faley, 1993). One goal of DEOMI is to keep the MEOCS up to date.

Barnes (1996) proposed that a revised MEOCS be modular. Each module should contain about five items and have an internal consistency (Cronbach's alpha) of at least .75 (Dansby, Johnson, McIntyre, & Truhon, 2001). One suggested revision is to make the items more neutral (i.e., replace references to "majority," "minority," "men," and "women" with more general terms "race" and "gender" and then use demographic information to determine the respondent's specific race and gender). To develop modules of five items, methods for shortening the MEOCS have included confirmatory factor analysis (McIntyre, 1999), cluster analysis (Truhon, 1999), and item response theory (IRT; Truhon, 2000).

*Item Response Theory*³

Throughout most of the history of psychology, test construction has been dependent upon classical test theory (CTT). CTT is based upon the assumption that a person's score on a test is the result of a true test score and error (i.e., $x_{ij} = t_{ij} + e_{ij}$). Certain rules of measurement follow from this assumption, including increasing the length of a test increases its reliability and that comparing test scores for different forms of a test is best when the forms are parallel. These rules have been challenged by IRT (Embretson, 1996).

IRT is a system of measurement using a model in which a person's ability or trait levels are dependent upon their responses to items as well as the qualities of these items. While not as old as CTT, IRT has a long history, probably beginning with the work of Lord (1952).

IRT involves examining performance at the item level. From the pattern of item responses an estimate of a person's latent ability or trait (θ) can be calculated, usually scaled with a mean of 0 and a standard deviation of 1. Items can be examined to determine their discrimination (a) and their difficulty (b). In this way, the relationship between a person's latent ability or trait and their performance on a set of items can be presented as an ogive curve called an item characteristic curve (ICC) or item characteristic function.

There are two important assumptions in IRT. First, the items that make up the test or scale must be *unidimensional*, i.e., they measure only one ability. The unidimensionality of a set of items is usually established by factor analysis or a similar technique⁴. Second, there is *local independence* among the responses, (i.e., once the latent ability is controlled for, there is no relationship among a person's responses to different items).

³ For a review of the basics of item response theory and a comparison with classical test theory, see Hambleton, Swaminathan, and Rogers (1991), and Embretson and Reise (2000).

⁴ Hattie (1984; 1985) has found many of the methods used for determining unidimensionality to be inappropriate or inadequate.

From the early work with dichotomous items, models of IRT of tests with polytomous responses, such as multiple-choice and Likert-type scales, ensued. The earliest of these is Samejima's (1969, 1997) graded response model. This model assumes that the categories of responses can be ordered, such as $i = 1, 2 \dots n$ where n is the highest level of response. It uses the formula below to calculate what are called category response functions (CRFs) for each choice of a particular item.

$$P(x = i) = ((1/(1+e^{-Da(\theta-b(i-J))}) - 1/(1+e^{-Da(\theta-b(i))}))$$

where

- $P(x = i)$ is the probability of a person giving response i ;
- e is a transcendental number equal to 2.718;
- D is a constant equal to 1.702, used to produce ogive curves;
- a is the discrimination of the item as represented by the slope of the ICC;
- θ is the latent ability or trait;
- b_i is the level of the trait needed to respond above threshold i with a probability 50% on the θ -axis.

From these CRFs for each item in a set of items, the boundary response functions (BRFs) can be calculated using the formula below.

$$P(\theta) = e^{-Da(\theta-b)}/(1+e^{-Da(\theta-b)})$$

Differential Item Functioning

IRT can be used to examine group differences. If members of two groups differ in their rate of responding to an item when there is the same latent trait (θ), differential item functioning (DIF) is said to occur (in other words, their a and b parameters are not equal). In CTT, this is often referred to item bias. However, in IRT, DIF and item bias are not considered identical terms. The researcher must consider whether the difference is relevant (i.e., reflect real differences between the groups) or irrelevant (i.e., is an artifact of the testing process) (see Camilli & Shepard, 1994; Zumbo, 1999). While item bias has negative connotations, DIF can be a positive finding in certain circumstances. For example, a revised version of an item may display DIF when compared to the original version of an item; yet an examination of the revised item shows improved psychometric qualities, such as better discrimination and between-category thresholds (i.e., higher a and b 's closer to zero)⁵.

An IRT analysis calculates the parameters (a , b and θ) for the items and group of participants tested. Consequently, with two groups these parameters are calculated separately and may not be measured on the same scale. Thus, after IRT has been done but before DIF can be determined, a method for linking the groups to the same metric or scale must be established. Stocking and Lord (1983) described such a procedure as shown below:

⁵ For a discussion of the use of IRT in test construction, see by Stark, Chernyshenko, Chuah, Lee, & Wadlington (2001).

$$\theta_j^* = A\theta_j + K$$

where

A is the slope,
 K is the intercept,
 θ_j is the j th examinee's trait level parameter in the metric of the current test, and
 θ_j^* is θ_j expressed in the target test metric

If the two tests are equal, then, A should equal 1 and K equal 0. When this transformation is performed, the current test's item parameters (the discrimination parameter a_i , and the difficulty parameter b_i) are also transformed in the following way:

$$a_i^* = \frac{a_i}{K},$$

and

$$b_i^* = Ab_i + K$$

where $i = 1, 2, \dots, n$ the number of test items.

Calculating these revised parameters usually involves an iterative process (e.g., Baker's [1992] EQUATE program). Once the parameters have been linked, DIF analysis can be performed.

Two recent procedures for calculating DIF are DFIT and SIBTEST. DFIT (Differential Functioning of Items and Tests; Raju, van der Linden, & Fleer [1995]) is a parametric procedure for comparing the test characteristic curves of the two groups, derived from ICCs, to determine whether DIF exists. In the DFIT procedure, if we let F stand for the focal group and R for the reference group, then

$$DIF_i = d_i^2$$

where

i is the item number, and
 d is $P_{iF}(\theta) - P_{iR}(\theta)$, (i.e., the difference between the probability functions $[P(\theta) = 1/(1+e^{-Da(\theta-b)})]$ for the focal and reference groups.

While DFIT is a parametric test, SIBTEST (Simultaneous Item Bias Test) is a nonparametric method developed as an extension of Shealy and Stout's (1993) multidimensional

item response theory. DIF frequently occurs because the items included in IRT are not unidimensional. SIBTEST allows small nuisance factors to be included in DIF analysis.

Chang, Mazzeo, and Roussos (1996) have developed an adaptation of SIBTEST for polytomous items (called POLY-SIBTEST). It is limited to uniform DIF (i.e., where the latent ability for the reference group on an item is always higher or lower than that for the focal group). Any analysis of an item with nonuniform DIF (where the curves cross over) will be somewhat doubtful if POLY-SIBTEST is used.⁶

For polytomous items DIF (d_i) can be defined as follows:

$$d_i = \bar{Y}_{Rk} - \bar{Y}_{Fk}, k = 0, \dots, n_H$$

where

$\bar{Y}_{Rk} - \bar{Y}_{Fk}$ is the difference in performance between the reference group and the focal group on the same observed matching test score; and

n_H is the maximum possible matching score.

SIBTEST calculates a β statistic

$$\hat{\beta} = \sum_{k=0}^{n_H} p_k d_k$$

where

p_k is the proportion of focal and reference group participants who attain a score of $X=k$.

The significance of β is determined by

$$B = \frac{\hat{\beta}}{\hat{\sigma}(\hat{\beta})}$$

where

B is normally distributed with $\bar{B} = 0$ and $s(B) = 1$; and

$$\hat{\sigma}(\hat{\beta}) = \left[\sum_{k=0}^{n_H} p_k^2 \left(\frac{\hat{\sigma}^2(Y|k, R)}{N_{Rk}} + \frac{\hat{\sigma}^2(Y|k, F)}{N_{Fk}} \right) \right]^{1/2}$$

⁶ There is a version of SIBTEST for nonuniform DIF but it only works for dichotomous items (Li & Stout, 1996).

where

$\hat{\sigma}^2(Y | k, g)$ is the variance of the item scores for participants in group g ($g = R$ or F) with score equal to k on the matching test.

Generally when $|\beta| > .100$, it is significant (Dorans & Holland, 1993).

A further advantage to these two programs is that they also measure differential test functioning (DTF). DTF is analogous to DIF; if sufficient items in a test display DIF, the test itself is said to DTF, sometimes called test bias⁷. In the DFIT format,

$$DTF = \sum_{i=1}^n d_i^2$$

The significance of DTF is determined by chi-square

$$\chi_N^2 = \frac{DTF}{\sigma_{DTF}^2}$$

where

N is the number of participants in the focal group.

In SIBTEST (Shealy & Stout, 1993), DTF occurs as the result of a combination of items showing DIF, i.e.,

$$B(\theta) = T_{SR}(\theta) - T_{SF}(\theta) = \sum_{i=1}^N T_{iR}(\theta) - \sum_{i=1}^N T_{iF}(\theta)$$

where

$T_{SR}(\theta)$ is the test characteristic curve for the reference group,
 $T_{SF}(\theta)$ is the test characteristic curve for the focal group,
 $T_{iR}(\theta)$ is the item characteristic curve for the reference group, and
 $T_{iF}(\theta)$ is the item characteristic curve for the focal group.

Applications to Equal Opportunity Research

There have been relatively few studies applying IRT or similar approaches to the study of equal opportunity (EO). One use of IRT has been to shorten scales. Stark, Chernyshenko, Lancaster, Drasgow, and Fitzgerald (2002) used Samejima's (1969) graded response model

⁷ The number of items with DIF does not necessarily determine whether a test displays DTF, especially if some items favor one group, while others favor the other group.

within the MULTILog program (Thissen, 1991) for the purpose of shortening the Sexual Experiences Questionnaire used by the Department of Defense (SEQ-DoD). Seven items from the 23-item scale could be eliminated while retaining strong psychometric qualities in the four subscales (Sexist Hostility [sexist behavior], Sexist Hostility [crude or offensive behavior], Unwanted Sexual Attention, and Sexual Coercion).

Similarly Truhon (2000) used IRT to shorten the five versions of the MEOCS: the Standard MEOCS, the MEOCS - Less Intensive Truncated Edition (LITE), the Senior Leader Equal Opportunity Climate Survey (SLEOCS), the MEOCS - Equal Employment Opportunity (EEO), and the Small Unit Equal Opportunity Climate Survey (SUEOCS). Eleven scales from the Standard MEOCS had acceptable psychometric qualities, while nine scales found only on one or more of the other four versions were possibly acceptable.

Both IRT and structural equation modeling (SEM) have been used to examine DIF⁸. Donovan and Drasgow (1999) applied Raju et al.'s (1995) DFIT program to examine sex differences in response to the SEQ-DoD. Significant DTF was found. Four items ("Treated you 'differently' because of your sex;" "Made offensive sexist remarks;" "Put you down or was condescending to you because of your sex;" and "Steered, leered, or ogled you in a way that made you feel uncomfortable") exhibited DIF. In all cases, the items discriminated equally well for men and women; the difference occurred because women showed a greater likelihood to endorse these items.

Schneider, Hitlan, and Radhakrishnan (2000) examined the structure of the newly developed Ethnic Harassment Experiences (EHE) scale. Using confirmatory factor analysis, they found the EHE consists of two scales: verbal harassment and exclusionary behavior. The number of factors, the loading of items on those factors, and the correlations between factors were equal for Hispanic and Anglo participants. However, Hispanics reported higher levels of ethnic jokes, while Anglos reported greater exclusion from social interactions.

Bergman, Palmieri, Drasgow, and Ormerod (2001) compared Whites', Blacks', Hispanics', Asian/Pacific Islanders', and Native American/Alaskan Natives' responses on a newly developed racial harassment and discrimination scale. Using confirmatory factor analysis, they found identical factor structure for each ethnic group. While non-Whites reported greater racial harassment and discrimination, the equivalent of DIF occurred for only one item ("Told stories or jokes which were racist or depicted your race/ethnicity negatively").

A pair of studies has made comparisons of IRT and SEM using the Satisfaction scale from the 1995 Armed Forces Sexual Harassment Survey (Edwards, Elig, Edwards, & Riemer, 1997). Collins, Raju, & Edwards (2000) found one item ("I am glad that I was assigned to this organization") displayed DIF when comparing men and women, as well as Blacks and Whites. One item ("To what extent does your chain of command provide you with the information you need to do your job?") displayed DIF when comparing men and women.

Raju, Laffitte, and Byrne (2002) found a lack of measurement equivalence for Blacks and Whites for the same item as Collins et al. (2000). They also found one item ("I have been taught

⁸ In SEM terms, a lack of measurement equivalence is the same thing as DIF (Raju, Laffitte, & Byrne, 2002).

valuable skills in the Service that I can use later in civilian jobs”) that displayed a lack of measurement equivalence using SEM but not with IRT. This may be due to the fact that SEM employs a linear model, while IRT a nonlinear model.

Purpose

Forty items from Standard MEOCS have been rewritten to express a more neutral quality. The purpose of the current study was to use IRT to compare the revised items with their original versions.

Method

Participants

Revised versions of MEOCS had been administered to 1,105 participants at the time of the current study. When cases with missing data were removed on the 40 items below, 1,040 participants remained. A random sample of 1,040 respondents to the Standard MEOCS with complete data for the comparable 40 items was selected for comparison.

Materials

Revised items have been established for eight scales from the Standard MEOCS: Sexual Harassment and Discrimination, Differential Command Behavior toward Minorities and Women, Positive Equal Opportunity (EO) Behavior, Racist/Sexist Behavior, Reverse Discrimination (Behavior), Discrimination against Minorities and Women, Reverse Discrimination (Attitude), and Attitudes toward Racial/Gender Separatism. Five items from each scale were chosen which previous research had shown to have good psychometric qualities (i.e., item-total correlations, reliability, and discriminability). Their means and standard deviations are presented in Table 1.

Table 1
Means and Standard Deviations from MEOCS and Comparable Revised Items

MEOCS Standard Item Number	Mean	Standard Deviation	MEOCS Revised Item Number	Mean	Standard Deviation
MEOCS 39	4.03	1.18	MEOCS-R 23*	4.33	1.01
MEOCS 43*	4.14	1.11	MEOCS-R 24	3.97	1.23
MEOCS 46	4.06	1.15	MEOCS-R 25*	4.39	0.93
MEOCS 47	4.13	1.16	MEOCS-R 26*	4.37	1.04
MEOCS 48	4.19	1.14	MEOCS-R 27	4.19	1.11
MEOCS 18	4.09	1.21	MEOCS-R 28*	4.30	1.02
MEOCS 23	4.20	1.12	MEOCS-R 29*	4.52	0.86
MEOCS 34	4.07	1.15	MEOCS-R 30*	4.46	0.88
MEOCS 38	4.05	1.22	MEOCS-R 31*	4.21	1.07
MEOCS 44*	4.10	1.20	MEOCS-R 32	3.78	1.32
MEOCS 5	2.39	1.35	MEOCS-R 37	2.42	1.39

MEOCS 7	2.16	1.31	MEOCS-R 39	2.10	1.26
MEOCS 29	2.19	1.32	MEOCS-R 40	2.15	1.26
MEOCS 35*	2.40	1.36	MEOCS-R 41	2.14	1.27
MEOCS 50	2.42	1.36	MEOCS-R 43	2.35	1.30
MEOCS 3*	3.80	1.26	MEOCS-R 93	3.64	1.27
MEOCS 9	4.16	1.16	MEOCS-R 94	4.13	1.10
MEOCS 12	4.03	1.15	MEOCS-R 95*	4.14	1.08
MEOCS 40	3.90	1.23	MEOCS-R 96*	4.40	0.98
MEOCS 42	3.78	1.28	MEOCS-R 97*	4.31	1.02
MEOCS 4	4.06	1.18	MEOCS-R 98*	4.17	1.07
MEOCS 17	3.99	1.20	MEOCS-R 99	4.00	1.15
MEOCS 22	3.86	1.29	MEOCS-R 100	3.80	1.26
MEOCS 33	4.20	1.11	MEOCS-R 101*	4.34	.95
MEOCS 45	3.95	1.23	MEOCS-R 102	3.99	1.19
RAPS 75	3.79	1.29	MEOCS-R 78*	4.01	1.24
RAPS 76	3.74	1.26	MEOCS-R 79*	3.90	1.19
RAPS 77	3.94	1.25	MEOCS-R 80	3.99	1.20
RAPS 85	3.74	1.26	MEOCS-R 81	3.79	1.26
RAPS 90*	3.94	1.23	MEOCS-R 82	3.69	1.32
RAPS 91	3.39	1.39	MEOCS-R 83	3.42	1.41
RAPS 93*	3.11	1.33	MEOCS-R 84	2.69	1.35
RAPS 96*	3.83	1.30	MEOCS-R 85	2.98	1.40
RAPS 99*	3.53	1.20	MEOCS-R 86	3.39	1.18
RAPS 100*	3.53	1.32	MEOCS-R 87	3.37	1.31
RAPS 80	4.35	1.07	MEOCS-R 88*	4.49	0.88
RAPS 82	4.03	1.15	MEOCS-R 89*	4.26	1.03
RAPS 87	4.35	1.06	MEOCS-R 90*	4.63	0.78
RAPS 88	4.39	1.00	MEOCS-R 91*	4.60	0.84
RAPS 92*	4.20	1.12	MEOCS-R 92	4.10	1.22

* $p < .05$

Nineteen of the 26 significantly different items display a more positive EO climate in the revised wording. (Two of the items where the Standard MEOCS have higher scores for Positive Equal Opportunity Behaviors items, which are reverse, coded compared to the other items). While that is not desirable from the standpoint of the sensitivity of the MEOCS (creating items whose means have less extreme values) and creating parallel forms, that parallelism is not necessary to create an alternate form of the MEOCS from the standpoint of IRT (Embretson, 1996).

Results

Thissen's (1991) MULTILOG program was used below to obtain difficulty and discriminability parameters (a and b 's). Because these parameters for the original version and the revised version were calculated separately, there needs to be a common metric. Baker's [1992] EQUATE program was used to equate the two versions. For each of the scales presented

below, the parameters from the revised form of the MEOCS were equated to those of the MEOCS Standard. The transformation constants (A and K) are also presented.

Following the transformation, DIF analyses were performed using Shealy and Stout's (1993) adapted for polytomous items (Chang et al., 1994).

Sexual Harassment and Discrimination

The discrimination and difficulty parameters for the original and revised versions of the Sexual Harassment and Discrimination scale are presented in Tables 2 and 3, respectively. Both versions show good discrimination and mostly negative discrimination parameters. These discrimination parameters suggest that the items discriminate better for those at the lower end of the scale. The marginal reliabilities for both scales are .78. The transformation constants for Table 3 are $A = 1.045$ and $K = .125$.

Table 2
Estimated Parameters for Sexual Harassment and Discrimination Items from the Standard MEOCS using Samejima's Graded Response Model

<u>Item Number</u>	<u>a</u>	<u>b₁</u>	<u>b₂</u>	<u>b₃</u>	<u>b₄</u>
MEOCS 39	1.18	-2.24	-1.67	-0.83	-0.06
MEOCS 43	1.30	-2.36	-1.75	-0.94	-0.16
MEOCS 46	1.30	-2.27	-1.62	-0.94	-0.05
MEOCS 47	1.44	-2.15	-1.59	-0.94	-0.23
MEOCS 48	1.62	-2.13	-1.62	-0.99	-0.32

- MEOCS 39 When a woman complained of sexual harassment to her superior, he told her, "You're being too sensitive."
- MEOCS 43 A woman was asked to take notes and provide refreshments at staff meetings (such duties were not part of her job assignment).
- MEOCS 46 A supervisor referred to female subordinates by their first names in public, while using titles for the male subordinates.
- MEOCS 47 The Commander/CO assigned an attractive woman to escort visiting male officials around because, "We need someone nice looking to show them around."
- MEOCS 48 A woman who complained of sexual harassment was not recommended for promotion.

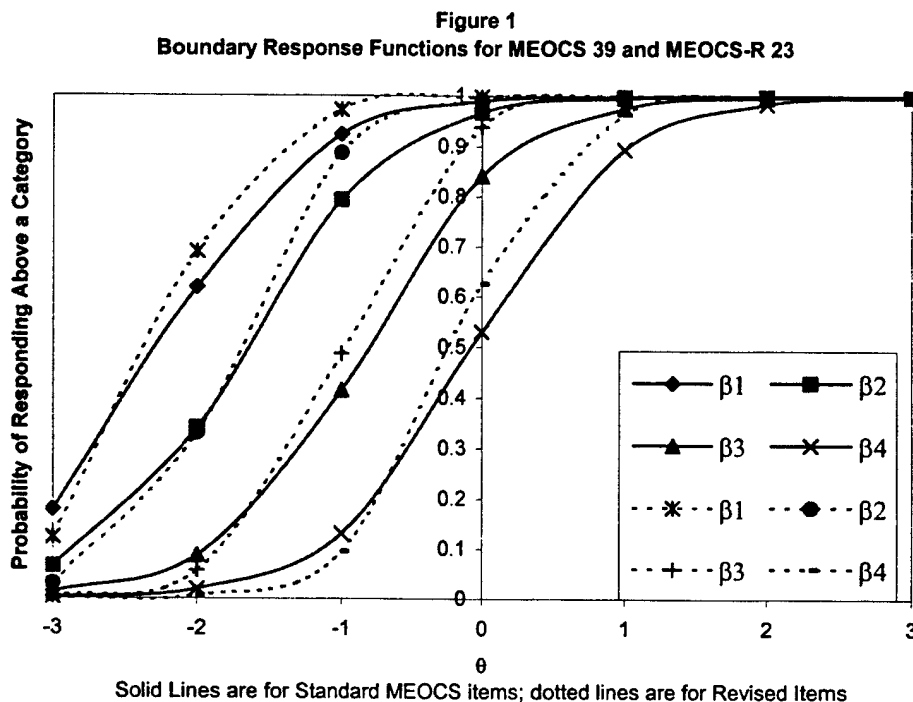
Table 3
Estimated Parameters for Sexual Harassment and Discrimination Items from the Standard MEOCS Revised using Samejima's Graded Response Model

<u>Item Number</u>	<u>a</u>	<u>b₁</u>	<u>b₂</u>	<u>b₃</u>	<u>b₄</u>
MEOCS 23	1.63	-2.29	-1.74	-0.98	-0.18

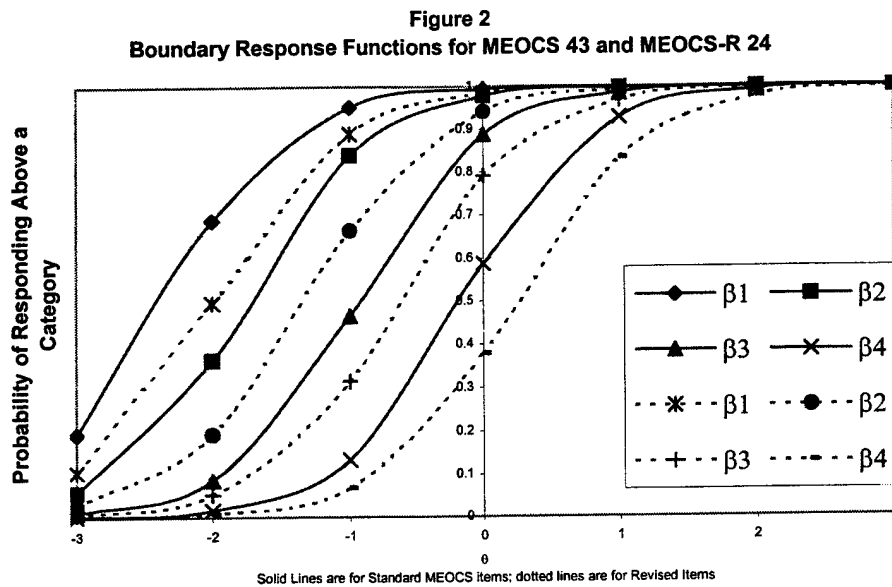
MEOCS 24	1.24	-2.00	-1.33	-0.63	0.23
MEOCS 25	1.52	-2.44	-1.99	-1.22	-0.19
MEOCS 26	1.43	-2.29	-1.71	-1.10	-0.38
MEOCS 27	1.50	-2.08	-1.55	-0.88	0.01

- MEOCS-R 23 When a person complained of sexual harassment, the supervisor said, "You're being too sensitive."
- MEOCS-R 24 A person was asked (because of their gender) to take notes and provide refreshments at staff meetings (such duties were not part of their job assignment).
- MEOCS-R 25 A supervisor referred to subordinates of the opposite gender by their first names in public, while using titles for subordinates of the same gender.
- MEOCS-R 26 The Commander/CO assigned an attractive person of the opposite gender to escort visiting officials because, "We need someone nice looking to show them around."
- MEOCS-R 27 A person who complained of sexual harassment was not recommended for promotion.

As can be seen in Figure 1, the BRFs for MEOCS 39 and MEOCS-R 23 are similar but with significant DIF ($\beta = -.203$, $p < .001$).



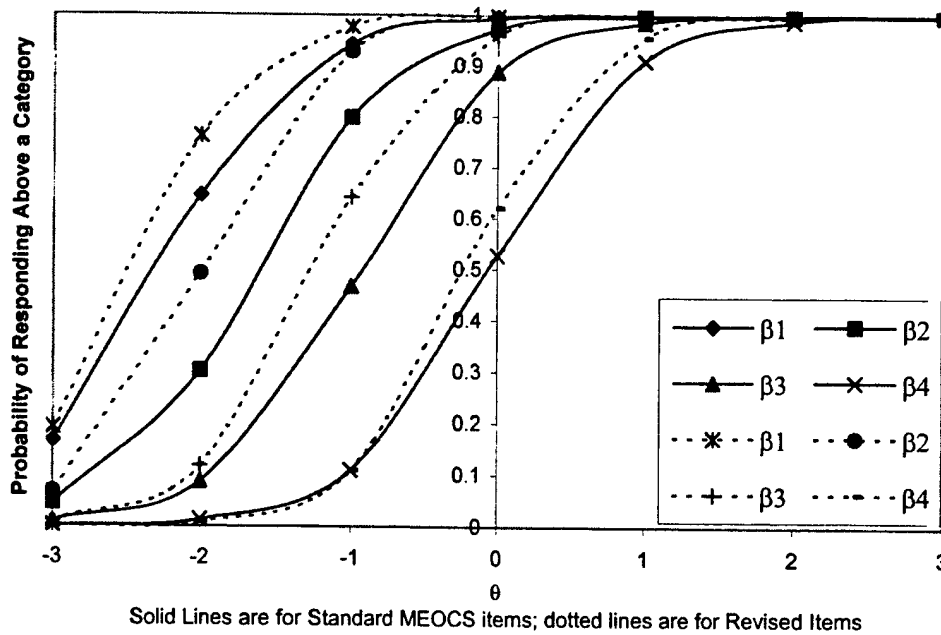
As can be seen in Figure 2, the BRFs for MEOCS 43 and MEOCS -R 24 show DIF ($\beta = .431, p < .001$). There is a uniform shift to the right with the revised item.⁹



As can be seen in Figure 3, the BRFs for MEOCS 46 and MEOCS-R 25 show DIF ($\beta = -.237, p < .001$). There is a uniform shift to the left with the revised item.

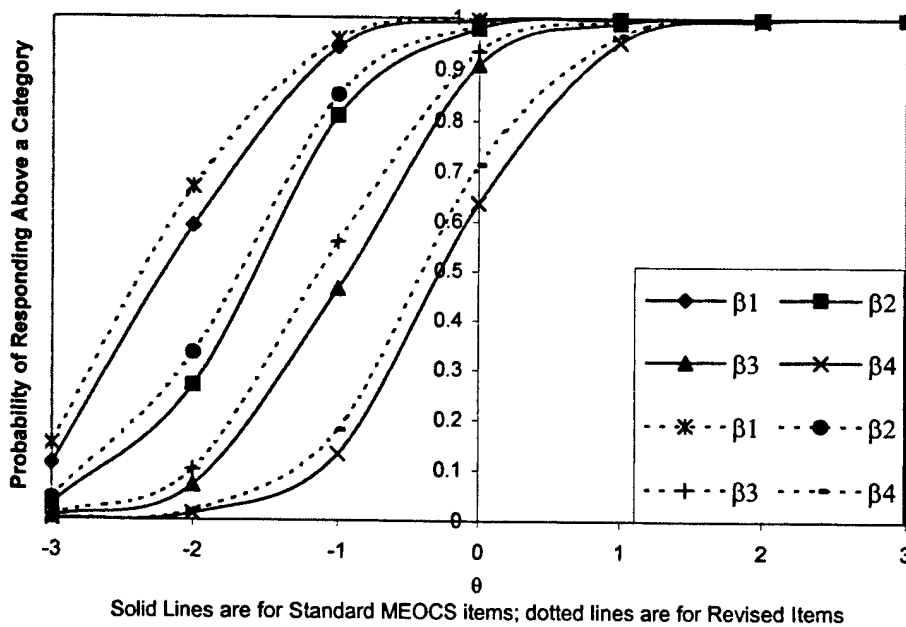
⁹ The significance of uniform and non-uniform shifts is presented in the Discussion section.

Figure 3
Boundary Response Functions for MEOCS 46 and MEOCS-R 25

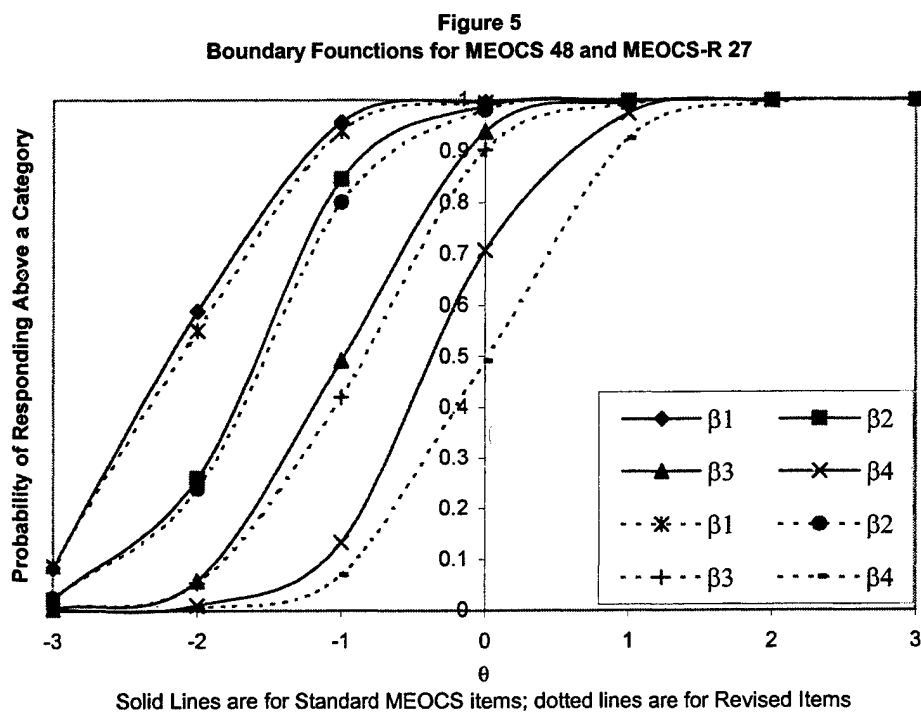


As can be seen in Figure 4, the BRFs for MEOCS 47 and MEOCS-R 26 are similar but with significant DIF ($\beta = -.134, p < .001$). There is a slight, but uniform, shift to the left.

Figure 4
Boundary Response Functions for MEOCS 47 and MEOCS-R 26



As can be seen in Figure 5, the BRFs for MEOCS 48 and MEOCS-R 27 are similar except for β_4 . There is significant DIF ($\beta = .215, p < .001$). In this case, there is a uniform shift to the right for the revised item.



Differential Command Behavior toward Minorities and Women

The discrimination and difficulty parameters for the original and revised versions of the Differential Command Behavior toward Minorities and Women scale are presented in Tables 4 and 5, respectively. Both versions show good discrimination and mostly negative discrimination parameters. These discrimination parameters suggest that the items discriminate better for those at the lower end of the scale. The marginal reliabilities for both scales are .78. The transformation constants for Table 5 are $A = 1.021$ and $K = .210$.

Table 4
Estimated Parameters for Differential Command Behavior toward Minorities and Women
Items from the Standard MEOCS using Samejima's Graded Response Model

<u>Item Number</u>	<u>a</u>	<u>b₁</u>	<u>b₂</u>	<u>b₃</u>	<u>b₄</u>
MEOCS 18	1.33	-2.01	-1.56	-0.90	-0.17
MEOCS 23	1.32	-2.21	-1.70	-1.04	-0.24
MEOCS 34	1.31	-2.17	-1.62	-0.85	-0.05
MEOCS 38	1.31	-2.01	-1.54	-0.84	-0.10
MEOCS 44	1.54	-1.93	-1.47	-0.87	-0.16

- MEOCS 18 A majority supervisor did not select a qualified minority subordinate for promotion.
- MEOCS 23 A minority member was assigned less desirable office space than a majority member.
- MEOCS 34 A motivational speech to a minority subordinate focused on the lack of opportunity elsewhere; to a majority subordinate, it focused on promotion.
- MEOCS 38 A qualified minority first-level supervisor was denied the opportunity for professional education by his/her supervisor. A majority first-level supervisor with the same qualifications was given the opportunity.
- MEOCS 44 A supervisor gave a minority subordinate a severe punishment for a minor infraction. A majority member who committed the same offense was given a less severe penalty.

Table 5

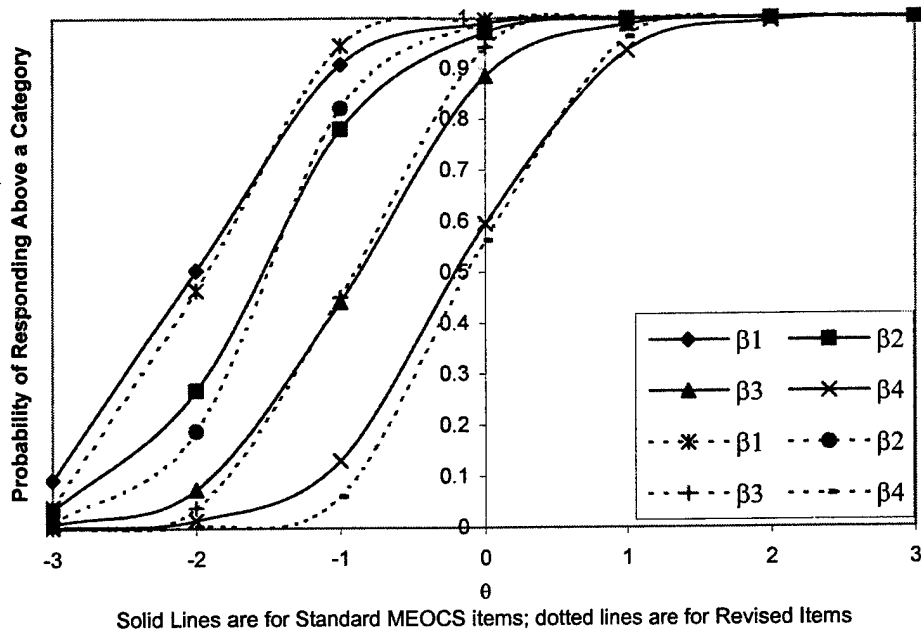
Estimated Parameters for Differential Command Behavior toward Minorities and Women Items from the Standard MEOCS Revised using Samejima's Graded Response Model

Item Number	a	b₁	b₂	b₃	b₄
MEOCS 28	1.75	-1.96	-1.51	-0.94	-0.08
MEOCS 29	2.09	-2.13	-1.74	-1.19	-0.38
MEOCS 30	2.06	-2.21	-1.77	-1.03	-0.28
MEOCS 31	1.20	-2.24	-1.64	-0.93	0.00
MEOCS 32	0.95	-1.77	-1.17	-0.41	0.48

- MEOCS-R 28 A supervisor did not select a qualified subordinate of another racial/ethnic background for promotion.
- MEOCS-R 29 A member of a particular racial/ethnic background was assigned less desirable office space than others in the organization.
- MEOCS-R 30 A motivational speech to a subordinate of a particular racial/ethnic background focused on the lack of opportunity elsewhere; to another subordinate, it focused on promotion.
- MEOCS-R 31 A qualified supervisor was denied the opportunity for professional education, but a supervisor of another racial/ethnic background with the same qualifications was given the opportunity.
- MEOCS-R32 A supervisor gave a subordinate of another racial/ethnic background a severe punishment for a minor infraction. A member with the same racial/ethnic background of the supervisor who committed the same offense was given a less severe penalty.

As can be seen in Figure 6, the BRFs for MEOCS 18 and MEOCS-R 28 are very similar with little DIF ($\beta = -.068$, $p = .075$).

Figure 6
Boundary Response Functions for MEOCS 18 and MEOCS-R 28



As can be seen in Figures 7 and 8, the BRFs for MEOCS 23 and MEOCS-R 29 and for MEOCS 34 and MEOCS-R 30 show non-uniform DIF ($\beta = -.223$ and $\beta = -.301$ respectively, p 's $< .001$). For these items, the original versions are higher at low θ , while the revised versions are higher at high θ .

Figure 7
Boundary Response Functions for MEOCS 23 and MEOCS-R 29

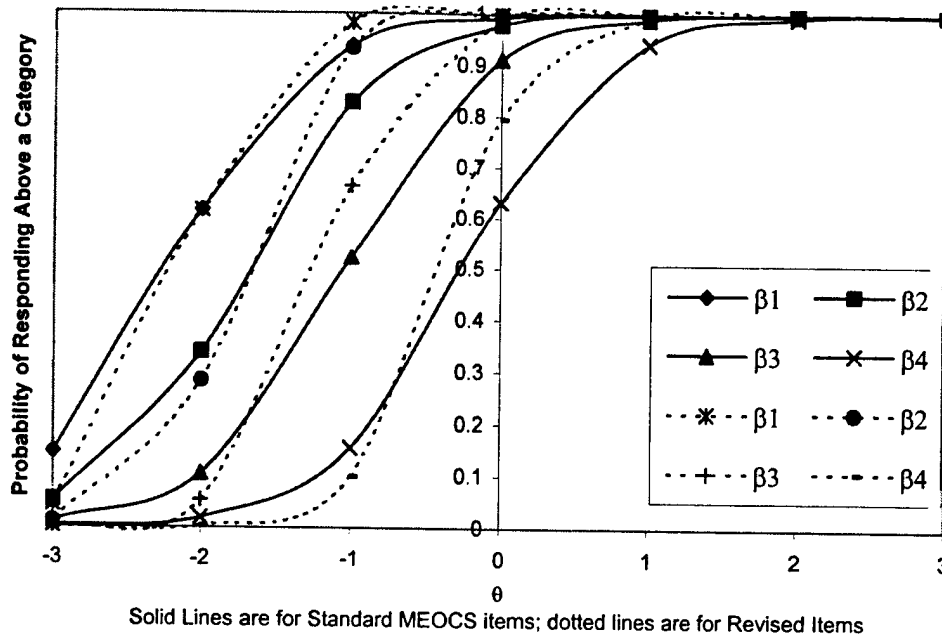
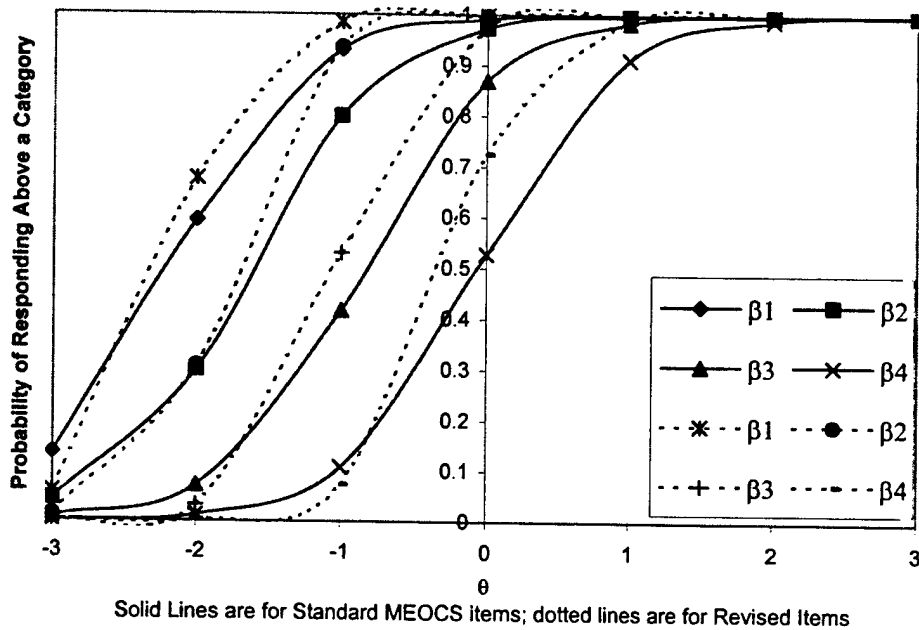
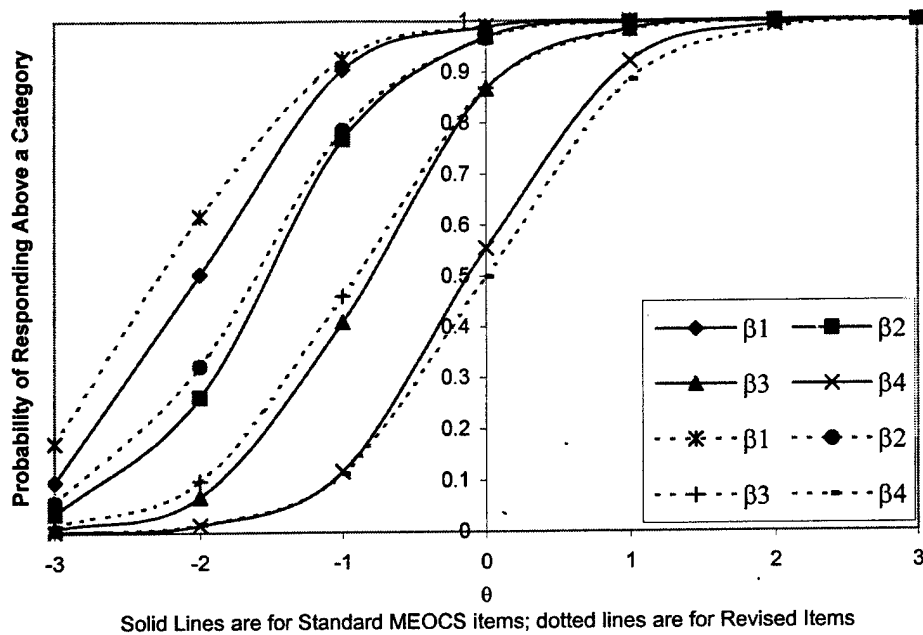


Figure 8
Boundary Response Functions for MEOCS 34 and MEOCS-R 30



As can be seen in Figure 9, the BRFs for MEOCS 38 and MEOCS-R 31 are very similar with little DIF ($\beta = -.022$, $p = .579$).

Figure 9
Boundary Response Functions for MEOCS 38 and MEOCS-R 31



As can be seen in Figure 10, the BRFs for MEOCS 44 and MEOCS -R 32 show DIF ($\beta = .690, p < .001$). There is a uniform shift to the right with the revised item.

Figure 10
Boundary Response Functions for MEOCS 44 and NEOCS-R 32

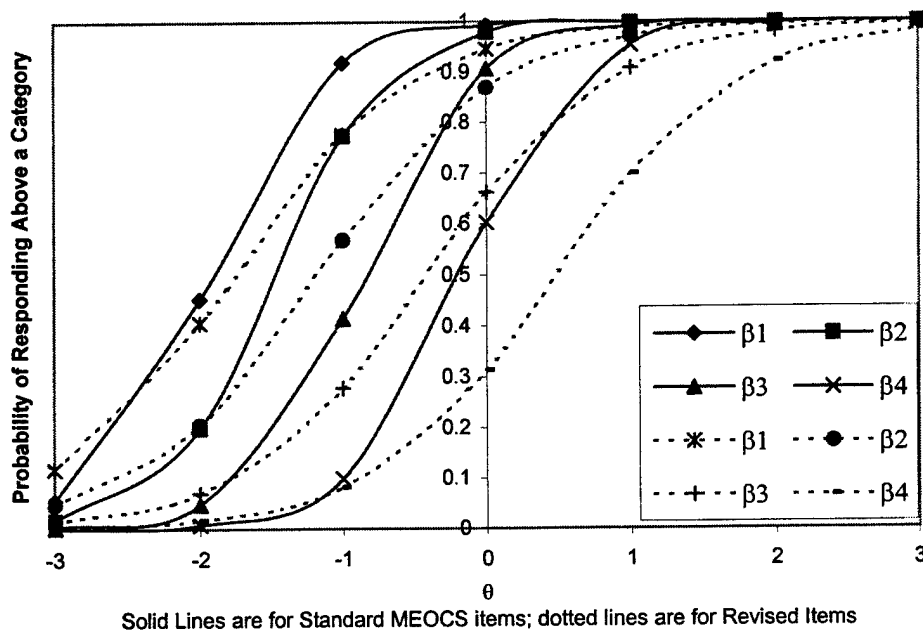
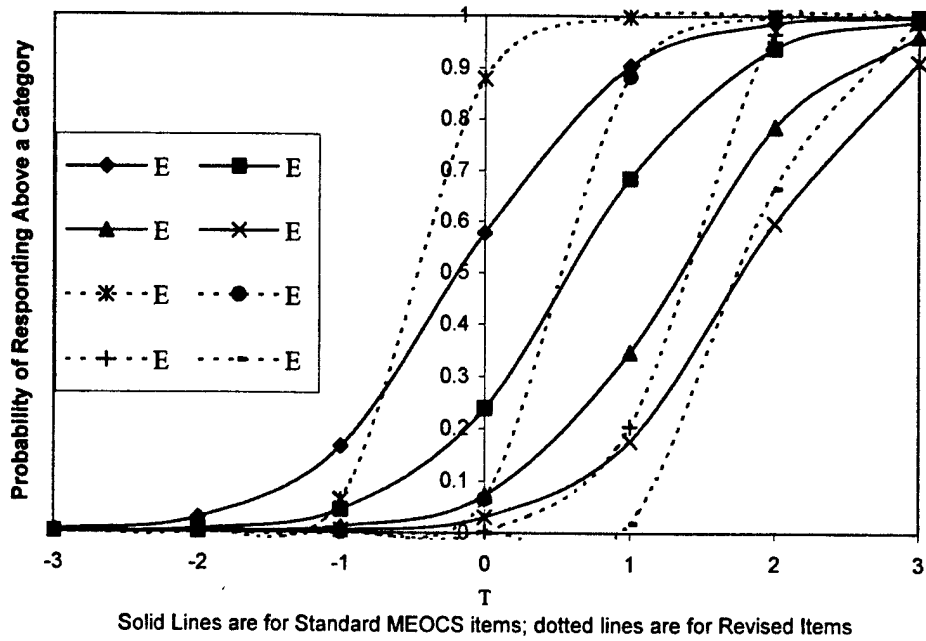
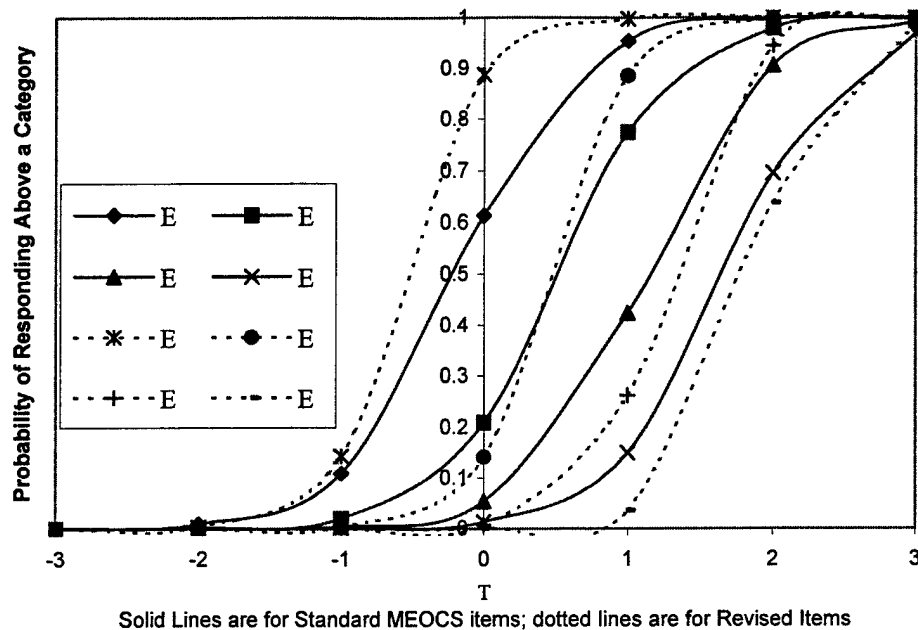


Figure 12
Boundary Response Functions for MEOCS 29 and MEOCS-R 40



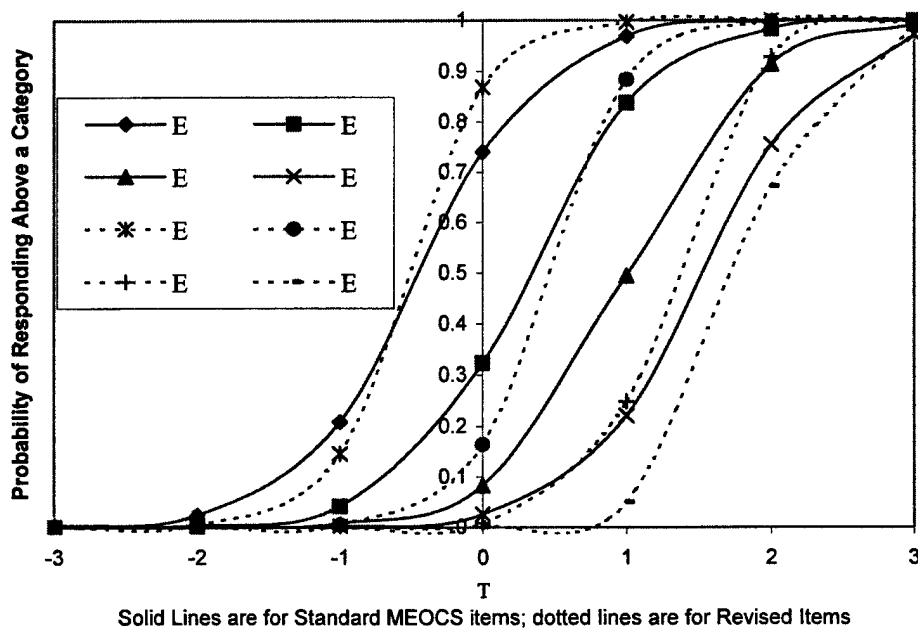
As can be seen in Figure 13, the BRFs for MEOCS 29 and MEOCS-R 40 show a mixed pattern ($E = .013$, $p = .715$). The lack of significance between MEOCS 29 and MEOCS-R 40 is probably due to the curves crossing over. For b_1 , there is a shift to the left. There is non-uniform DIF for b_2 and b_3 . For b_4 , there is a shift to the right.

Figure 13
Boundary Response Functions for MEOCS 29 and MEOCS-R 40

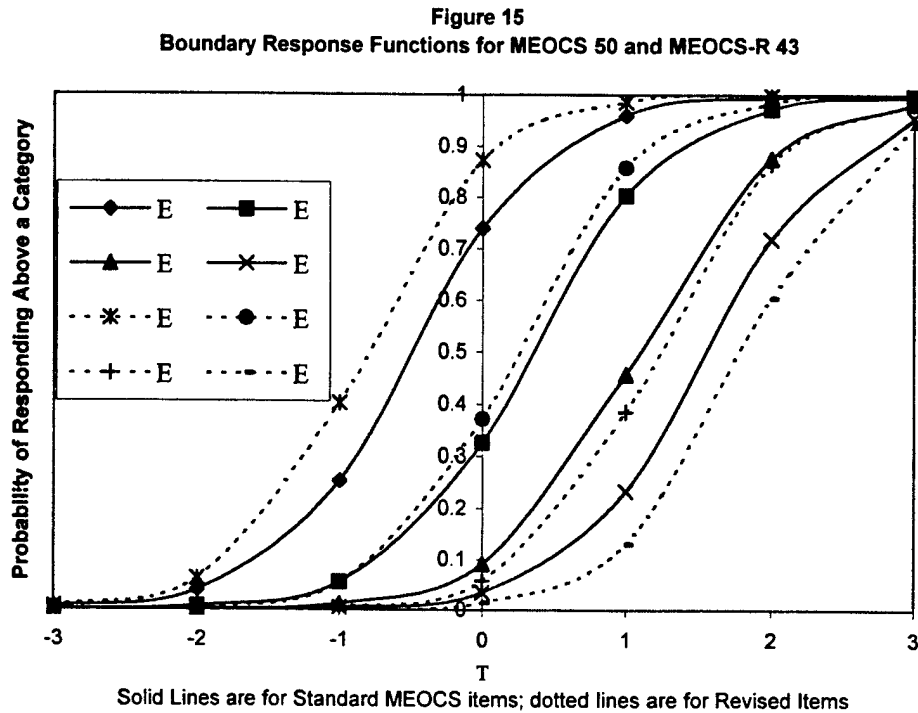


As can be seen in Figure 14, the BRFs for MEOCS 35 and MEOCS-R 41 show a mixed pattern ($E = .260, p < .001$). There is non-uniform DIF for b_1 and b_2 . For b_3 and b_4 , there is a shift to the right.

Figure 14
Boundary Response Functions for MEOCS 35 and MEOCS-R 41



As can be seen in Figure 15, the BRFs for MEOCS 50 and MEOCS-R 43 show a mixed pattern ($E = -.072$, $p = .090$). For b_1 and b_2 , there is a shift to the left; for b_3 and b_4 , there is a shift to the right.



Racist/Sexist Behaviors

The discrimination and difficulty parameters for the original and revised versions of the Racist/Sexist Behavior scale are presented in Tables 8 and 9 respectively. Both versions show good discrimination and mostly negative discrimination parameters. These discrimination parameters suggest that the items discriminate better for those at the lower end of the scale. The marginal reliabilities for both scales are .81 and .82, respectively. The transformation constants for Table 9 are $A = 1.201$ and $K = .409$.

Table 8
Estimated Parameters for Racist/Sexist Behavior Items from the Standard MEOCS
using Samejima's Graded Response Model

<u>Item</u> <u>Number</u>	<u>a</u>	<u>b₁</u>	<u>b₂</u>	<u>b₃</u>	<u>b₄</u>
MEOCS 3	0.84	-2.28	-1.65	-0.75	0.37
MEOCS 9	0.91	-2.69	-1.92	-1.16	-0.32
MEOCS 12	1.15	-2.48	-1.66	-0.89	0.00
MEOCS 40	1.76	-1.83	-1.36	-0.64	0.10
MEOCS 42	1.78	-1.74	-1.17	-0.57	0.23

- MEOCS 3 A majority person told several jokes about minorities.
MEOCS 9 A majority member in your organization directed a racial slur at a member of another organization.
MEOCS 12 A group of majority and minority personnel made reference to an ethnic group other than their own using insulting ethnic names.
MEOCS 40 Offensive racial/ethnic names were frequently heard.
MEOCS 42 Racial/ethnic jokes were frequently heard.

Table 9

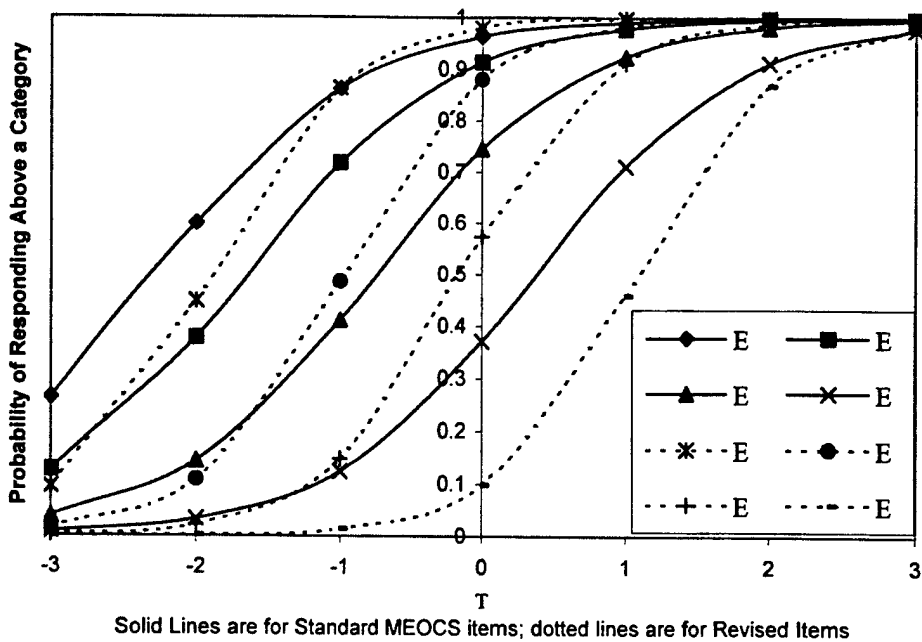
**Estimated Parameters for Racist/Sexist Behavior Items from the Standard MEOCS
Revised using Samejima's Graded Response Model**

<u>Item Number</u>	<u>a</u>	<u>b₁</u>	<u>b₂</u>	<u>b₃</u>	<u>b₄</u>
MEOCS 93	1.20	-1.90	-0.97	-0.14	1.08
MEOCS 94	1.58	-2.31	-1.57	-0.74	0.36
MEOCS 95	1.79	-2.29	-1.62	-0.72	0.36
MEOCS 96	1.91	-2.51	-1.82	-1.20	-0.16
MEOCS 97	1.44	-2.57	-1.87	-1.10	0.02

- MEOCS-R 93 A person of a particular racial/ethnic background told several jokes about people of a different racial/ethnic background.
MEOCS-R 94 A member in your organization directed a racial slur at a member of another organization.
MEOCS-R 95 A group of personnel of different racial/ethnic backgrounds made reference to an ethnic group other than their own using insulting ethnic names.
MEOCS-R 96 Offensive racial/ethnic names were frequently heard.
MEOCS-R 97 Racial/ethnic jokes were frequently heard.

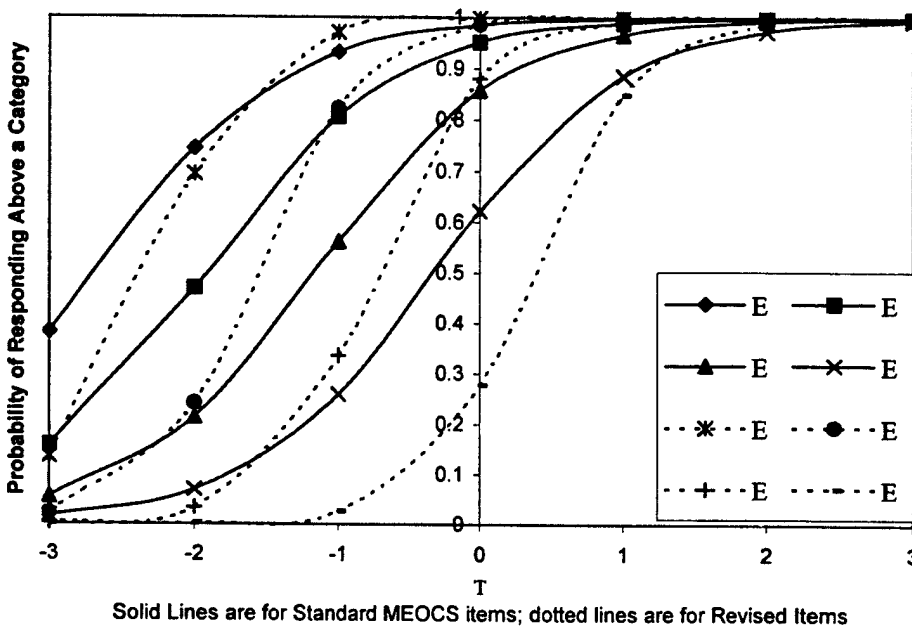
As can be seen in Figure 16, the BRFs for MEOCS 3 and MEOCS-R 93 show DIF ($E = .583$, $p < .001$). There is a shift to the right with the revised item.

Figure 16
Boundary Response Functions for MEOCS 3 and MEOCS-R 93

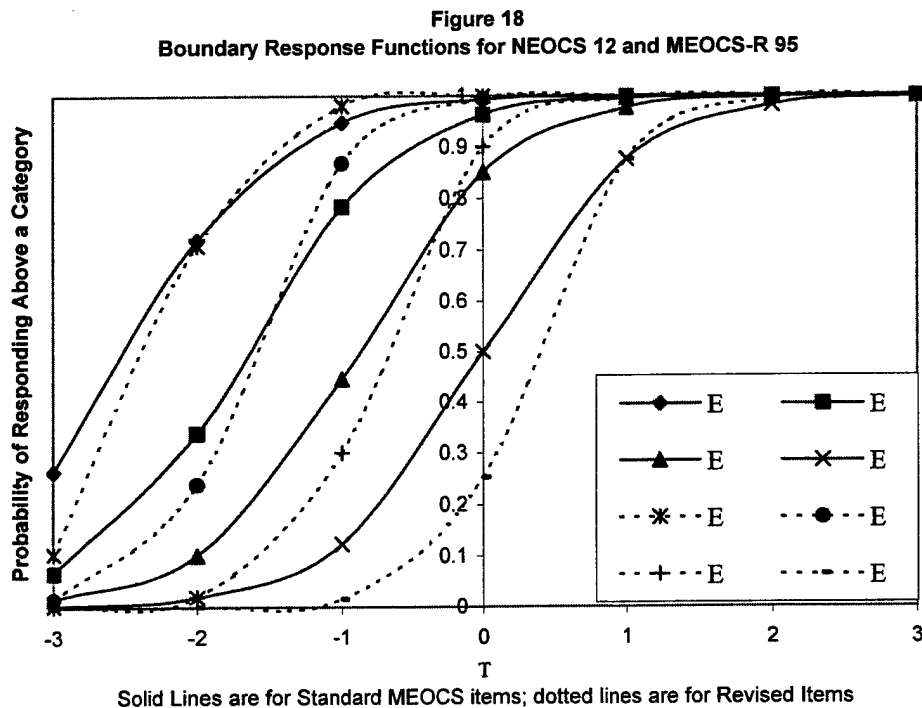


As can be seen in Figure 17, the BRFs for MEOCS 9 and MEOCS-R 94 show nonuniform DIF ($E = .255$, $p < .001$). For this item, the original version is higher at low T , while the revised version is higher at very high T .

Figure 17
Boundary Response Functions for MEOCS 9 and MEOCS-R 94



As can be seen in Figure 18, the BRFs for MEOCS 12 and MEOCS-R 95 show non-uniform DIF ($E = .096$, $p = .003$). For this item, the revised version is higher at low T , while the original version is higher at very high T .



As can be seen in Figures 19 and 20, the BRFs for MEOCS 40 and MEOCS-R 96 and for MEOCS 42 and MEOCS-R 97 show DIF ($E = -.445$ and $E = -.443$, respectively, p 's $< .001$). There is a shift to the left with the revised items.

Figure 19
Boundary Response Functions for MEOCS 40 and MEOCS-R 96

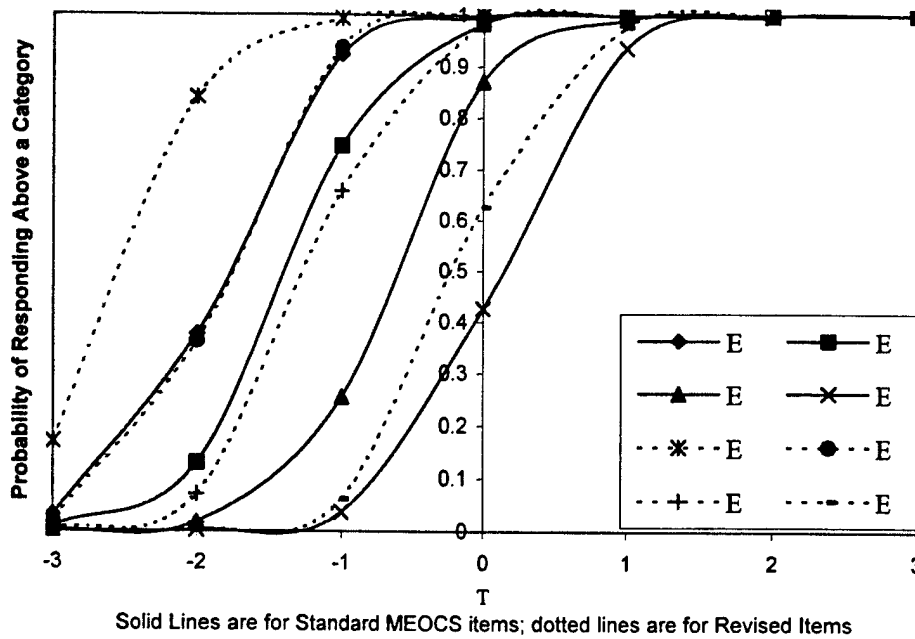
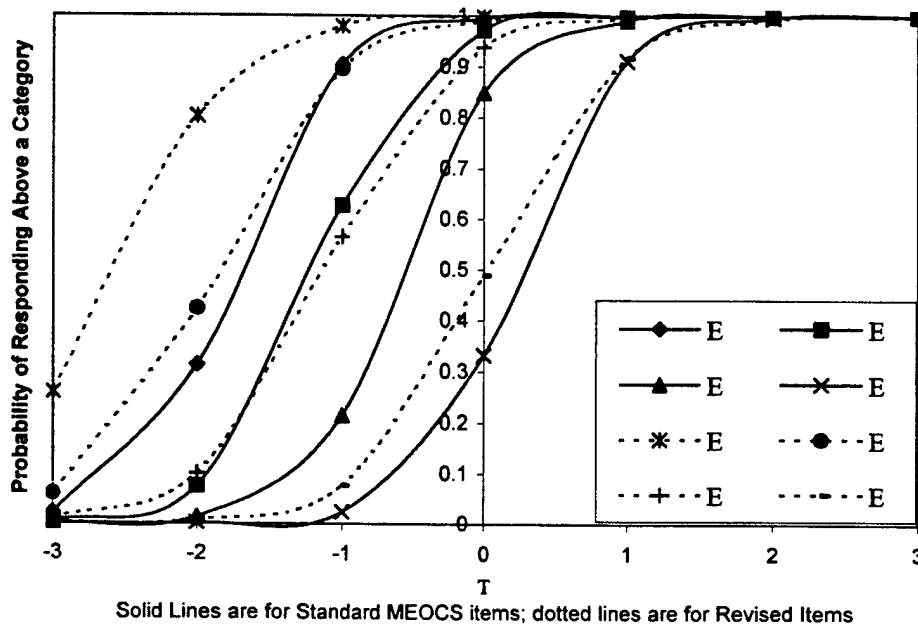


Figure 20
Boundary Response Functions for MEOCS 42 and MEOCS-R 97



Reverse Discrimination (Behavior)

The discrimination and difficulty parameters for the original and revised versions of the Reverse Discrimination (Behavior) scale are presented in Tables 10 and 11,

respectively. Both versions show good discrimination and mostly negative discrimination parameters. These discrimination parameters suggest that the items discriminate better for those at the lower end of the scale. The marginal reliabilities for both scales are .76 and .82, respectively. The transformation constants for Table 11 A = 1.267 and K = .420.

Table 10
Estimated Parameters for Reverse Discrimination (Behavior) Items from the Standard MEOCS using Samejima's Graded Response Model

<u>Item Number</u>	<u>a</u>	<u>b₁</u>	<u>b₂</u>	<u>b₃</u>	<u>b₄</u>
MEOCS 4	0.76	-2.79	-2.00	-1.06	-0.06
MEOCS 17	1.32	-2.04	-1.45	-0.72	0.05
MEOCS 22	1.19	-1.91	-1.32	-0.58	0.12
MEOCS 33	0.99	-2.56	-1.96	-1.04	-0.26
MEOCS 45	1.07	-2.23	-1.46	-0.69	0.06

- MEOCS 4 The Commander/CO did not appoint a qualified majority in a key position, but instead appointed a less qualified minority.
- MEOCS 17 A minority man was selected for a prestigious assignment over a majority man who was equally, if not slightly better, qualified.
- MEOCS 22 A minority woman was selected to receive an award for an outstanding act even though her peers as did not perceive her being as qualified as her nearest competitor, a majority man.
- MEOCS 33 A majority and a minority person turned in similar pieces of equipment with similar problems. The minority person was given a new issue; the majority member's equipment was sent to maintenance for repair.
- MEOCS 45 A better qualified man was not picked for a good additional duty assignment because the Commander/CO said it would look better for equal opportunity to have a woman take this duty.

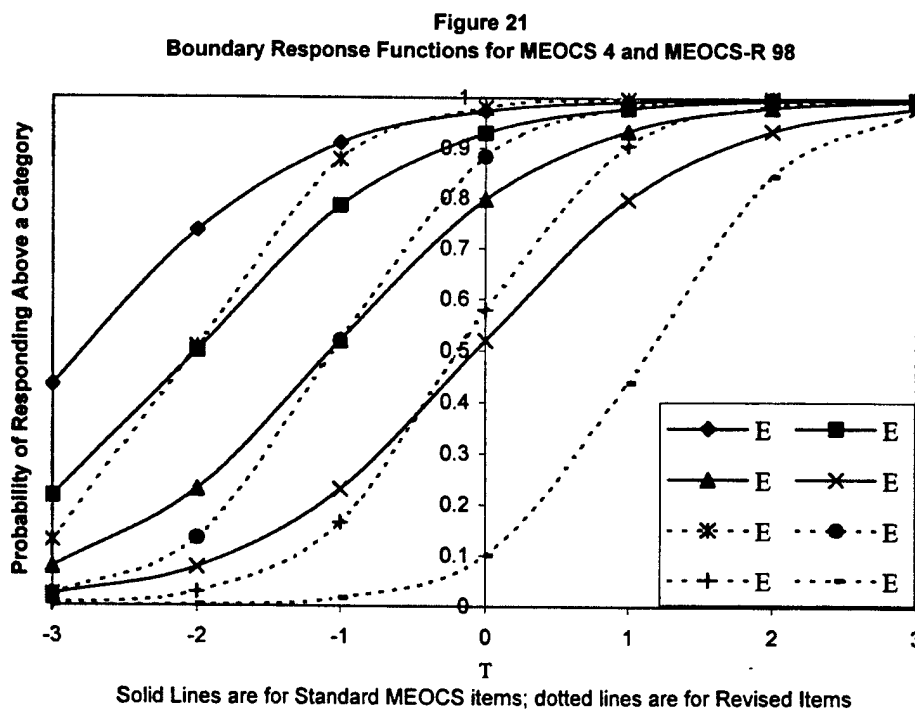
Table 11
Estimated Parameters for Reverse Discrimination (Behavior) Items from the Standard MEOCS Revised using Samejima's Graded Response Model

<u>Item Number</u>	<u>a</u>	<u>b₁</u>	<u>b₂</u>	<u>b₃</u>	<u>b₄</u>
MEOCS 98	1.14	-2.01	-1.04	-0.16	1.13
MEOCS 99	1.49	-2.44	-1.67	-0.80	0.36
MEOCS 100	1.69	-2.43	-1.72	-0.78	0.36
MEOCS 101	1.81	-2.66	-1.94	-1.28	-0.18
MEOCS 102	1.37	-2.72	-1.99	-1.18	0.01

- MEOCS-R 98 The Commander/CO did not appoint a qualified member to a key position, but instead appointed a less qualified member of a different

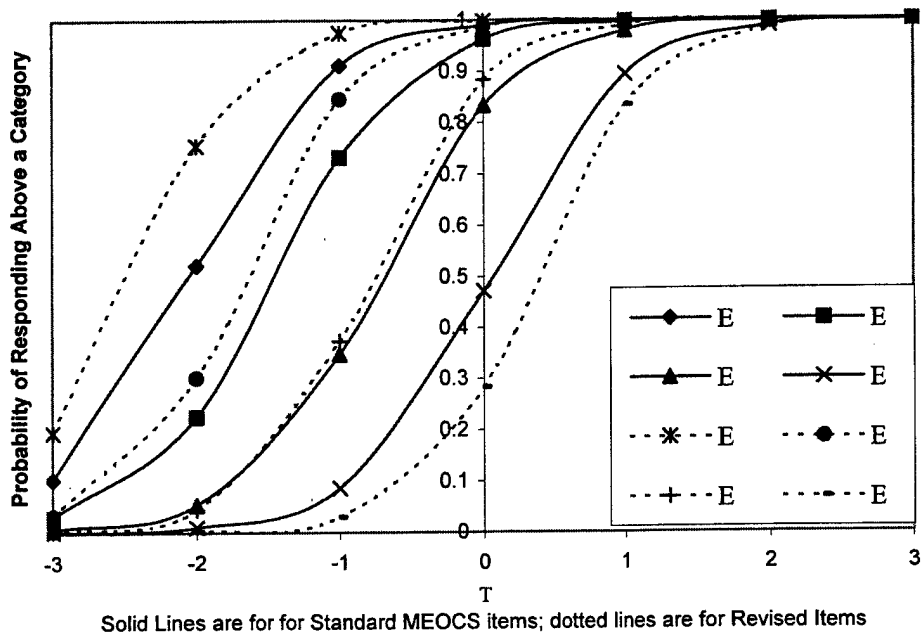
- racial/ethnic group.
- MEOCS-R 99 A person of one race/ethnicity was selected for a prestigious assignment over a person of another race/ethnicity who was equally, if not slightly better, qualified.
- MEOCS-R 100 A person of a specific gender and race/ethnicity was selected to receive an award for an outstanding act even though the person was not perceived by peers as being as qualified as the nearest competitor, a person of the opposite gender and of a different race/ethnicity.
- MEOCS-R 101 Two people of differing race/ethnicity/gender turned in similar pieces of equipment with similar problems. One person was given a new issue; the other's equipment was sent to maintenance for repair.
- MEOCS-R 102 A better qualified person of a specific gender was not picked for a good additional duty assignment because the Commander/CO said it would look better for equal opportunity to have a person of that gender take this duty.

As can be seen in Figure 21, the BRFs for MEOCS 4 and MEOCS-R 98 show a little DIF ($E = -.086$, $p = .031$). There is a uniform shift to the right with the revised item.



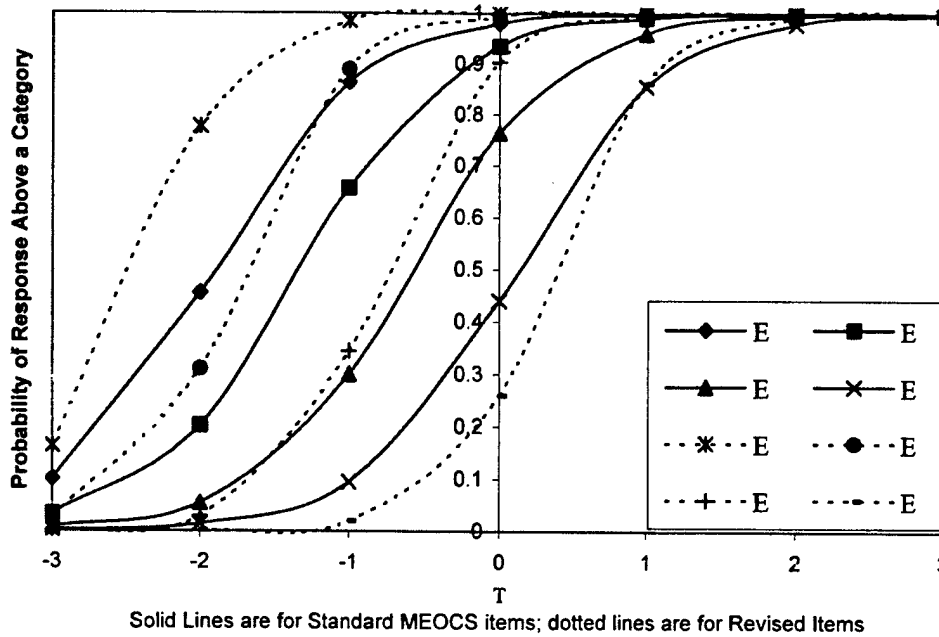
As can be seen in Figure 22, the BRFs for MEOCS 17 and MEOCS-R 99 show a mixed pattern ($E = .061$, $p = .092$). The lack of significance between MEOCS 17 and MEOCS-R 99 is probably due to the different patterns for the curve pairs. For b_1 and b_2 , there is a shift to the left; for b_3 , the curves are nearly identical, and for b_4 , there is a shift to the right.

Figure 22
Boundary Response Functions for MEOCS 17 and MEOCS-R 99



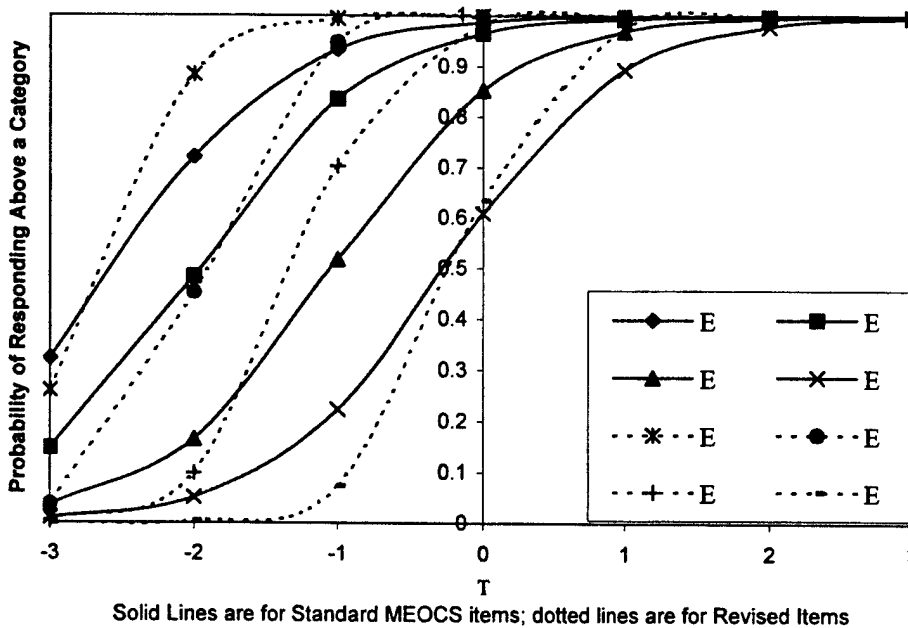
As can be seen in Figure 23, the BRFs for MEOCS 22 and MEOCS-R 100 show a mixed pattern ($E = .264$, $p < .001$). For b_1 and b_2 , there is a shift to the left. There is non-uniform DIF for b_3 and b_4 . For this item, the original version is higher at low T , while the revised version is higher at high T .

Figure 23
Boundary Response Functions for MEOCS 22 and MEOCS-R 100

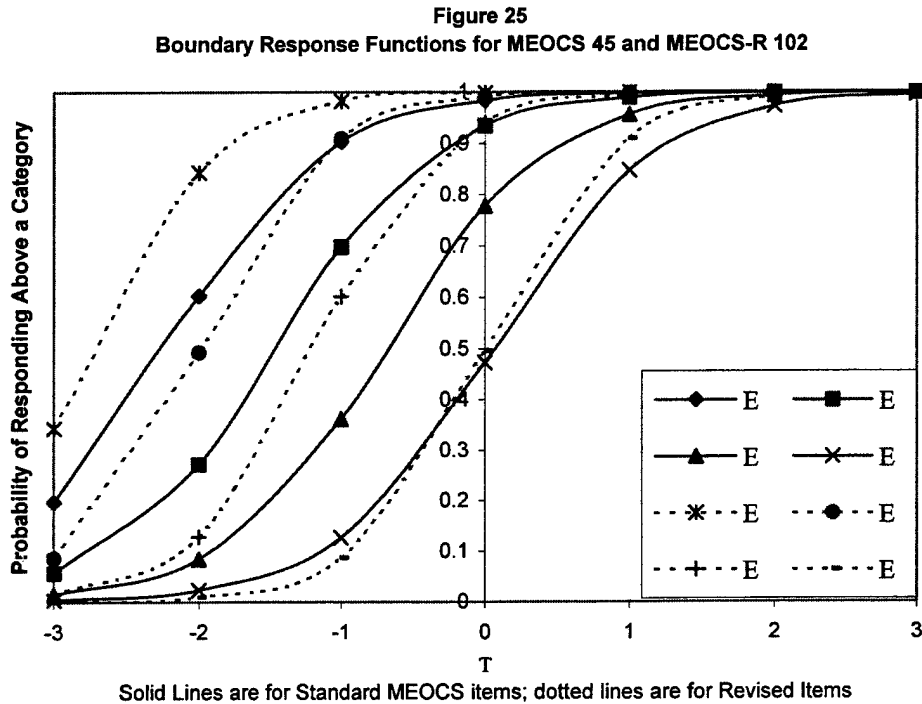


As can be seen in Figure 24, the BRFs for MEOCS 33 and MEOCS-R 101 show non-uniform DIF ($E = -.114$, $p = .002$). For this item, the original version is higher at low T , while the revised version is higher at high T .

Figure 24
Boundary Response Function for MEOCS 33 and MEOCS-R 101



As can be seen in Figure 25, the BRFs for MEOCS 45 and MEOCS-R 102 show a mixed pattern ($E = .067$, $p = .095$). The lack of significance between MEOCS 45 and MEOCS-R 102 is probably due to the different patterns for the curve pairs. For b_1 , b_2 , and b_3 there is a shift to the left. There is non-uniform DIF for b_4 . For these two BRFs, the original versions are higher at low T , while the revised versions are higher at high T .



Discrimination against Minorities and Women

The discrimination and difficulty parameters for the original and revised versions of the Discrimination against Minorities and Women scale are presented in Tables 12 and 13, respectively. Both versions show good discrimination and mostly negative discrimination parameters. These discrimination parameters suggest that the items discriminate better for those at the lower end of the scale. The marginal reliabilities for both scales are .84 and .85 respectively. The transformation constants for Table 13 $A = 1.010$ and $K = .061$.

Table 12
Estimated Parameters for Discrimination against Minorities and Women Items
from the Standard MEOCS using Samejima's Graded Response Model

<u>Item</u> <u>Number</u>	<u>a</u>	<u>b₁</u>	<u>b₂</u>	<u>b₃</u>	<u>b₄</u>
RAPS 75	2.79	-1.66	-1.02	-0.35	0.27
RAPS 76	3.18	-1.67	-0.96	-0.29	0.38
RAPS 77	3.04	-1.72	-1.18	-0.45	0.08
RAPS 85	1.89	-1.95	-1.13	-0.31	0.41
RAPS 90	2.40	-1.94	-1.26	-0.46	0.11

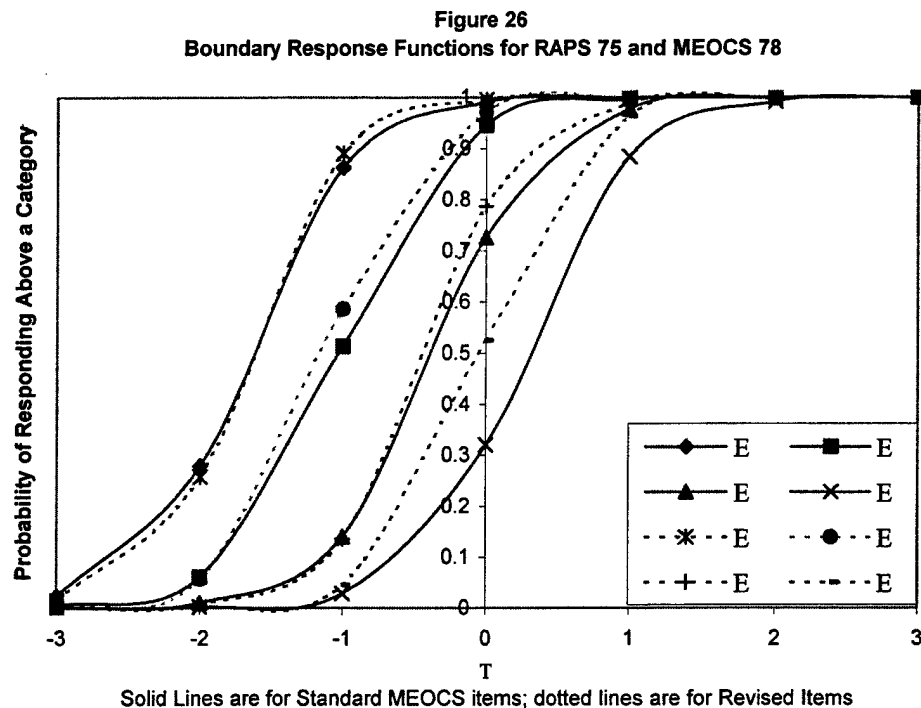
- RAPS 75 More severe punishments are given out to minority as compared to majority offenders for the same types of offenses.
- RAPS 76 Majority supervisors in charge of minority supervisors doubt the minorities' abilities.
- RAPS 77 Minorities get more extra work details than majority members.
- RAPS 85 Majority members assume that minorities commit every crime that occurs, such as thefts in living quarters.
- RAPS 90 Majority members get away with breaking rules that result in punishment for minorities.

Table 13
Estimated Parameters for Discrimination against Minorities and Women Items
from the Standard MEOCS Revised using Samejima's Graded Response Model

<u>Item</u> <u>Number</u>	<u>a</u>	<u>b₁</u>	<u>b₂</u>	<u>b₃</u>	<u>b₄</u>
MEOCS 78	1.86	-1.66	-1.11	-0.41	-0.03
MEOCS 79	1.76	-1.90	-1.14	-0.28	0.24
MEOCS 80	2.12	-1.73	-1.12	-0.40	0.09
MEOCS 81	1.19	-2.10	-1.04	-0.26	0.33
MEOCS 82	1.18	-1.91	-0.88	-0.20	0.40

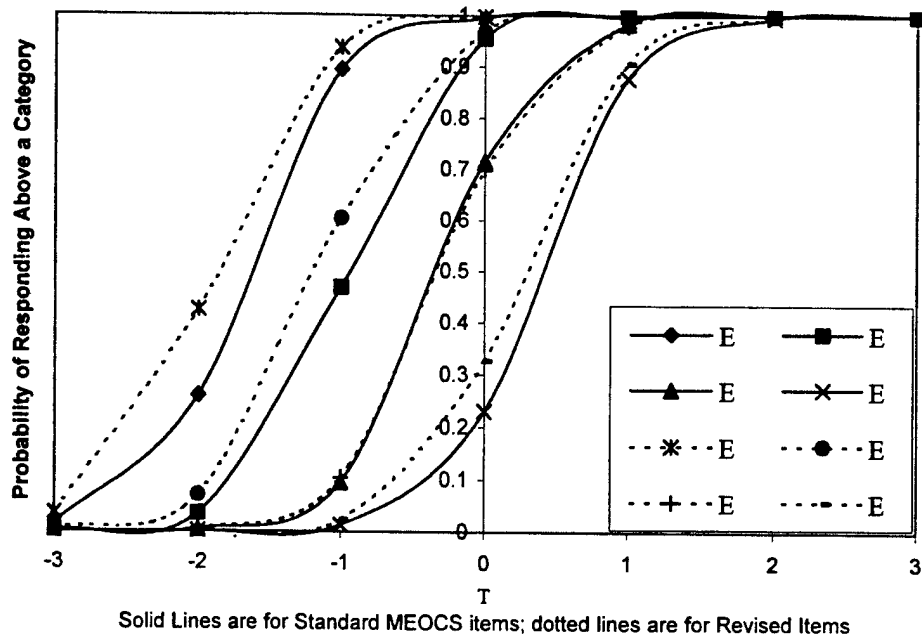
- MEOCS-R 78 More severe punishments are given out to members of some gender, racial and ethnic groups than members of other groups for the same types of offenses.
- MEOCS-R 79 Second-level supervisors doubt the abilities of first-level supervisors from some gender, racial, and ethnic groups.
- MEOCS-R 80 Members of some gender, racial, and ethnic groups get more extra work details than members of other groups.
- MEOCS-R 81 Some individuals assume that persons of certain racial or ethnic groups commit every crime that occurs, such as thefts in living quarters.
- MEOCS-R 82 Some racial/ethnic groups get away with breaking rules that result in punishment for others.

As can be seen in Figure 26, the BRFs for RAPS 75 and MEOCS-R 78 show a mixed pattern ($E = .224$, $p < .001$). For b_1 , b_2 , and b_3 the curves are nearly identical, and for b_4 , there is a shift to the left.



As can be seen in Figure 27, the BRFs for RAPS 76 and MEOCS-R 79 show a mixed pattern ($E = -.142$, $p < .001$). For b_1 , b_2 , and b_4 , there is a shift to the left; for b_3 , the curves are nearly identical.

Figure 27
Boundary Response Functions for RAPS 76 and MEOCS-R 79



As can be seen in Figures 28 and 29, the BRFs for RAPS 77 and MEOCS-R 80 and for RAPS 85 and MEOCS-R 81 are very similar with little DIF ($E = -.005$, $p = .881$, and $E = -.007$, $p = .879$, respectively).

Figure 28
Category Response Functions for RAPS 77 and MEOCS-R 80

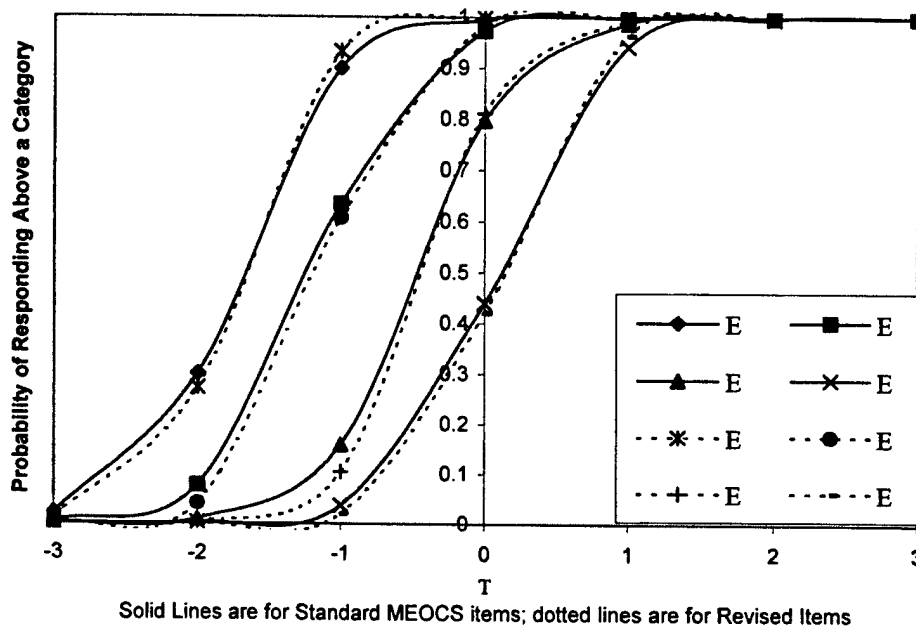
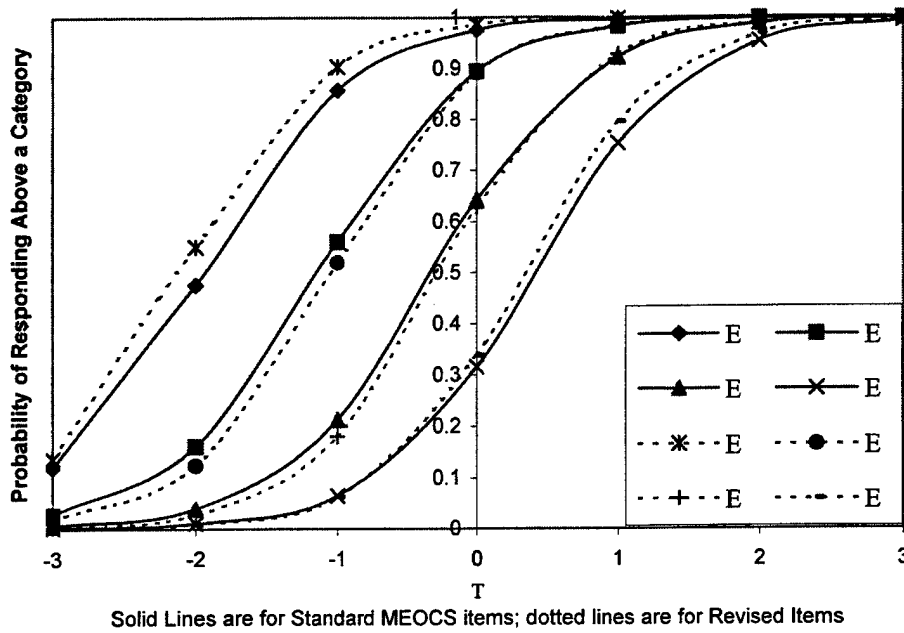
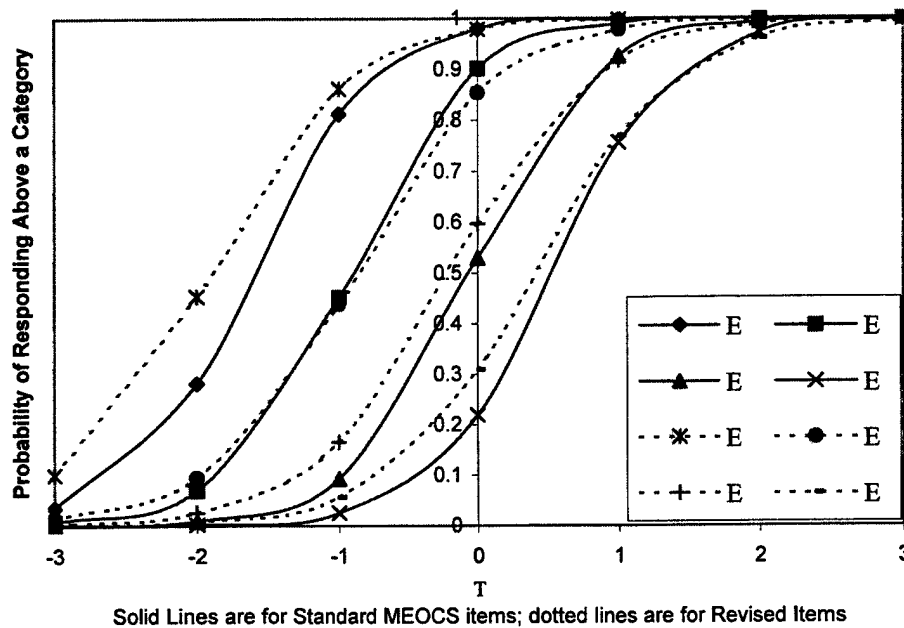


Figure 29
Boundary Response Functions for RAPS 85 and MEOCS-R 81



As can be seen in Figure 30, the BRFs for RAPS 90 and MEOCS-R 82 show a mixed pattern ($E = .369$, $p < .001$). For b_1 , b_3 , and b_4 there is a shift to the left, and for b_4 , the curves are nearly identical.

Figure 30
Boundary Response Functions for RAPS 90 and MEOCS 82



Reverse Discrimination (Attitudes)

The discrimination and difficulty parameters for the original and revised versions of the Reverse Discrimination (Attitudes) scale are presented in Tables 14 and 15, respectively. Both versions show good discrimination and both positive and negative discrimination parameters. These discrimination parameters suggest that the items discriminate about equally well at both ends of the scale. The marginal reliabilities for both scales are .81 and .86, respectively. The transformation constants for Table 15 A = 1.157 and K = -.243.

Table 14
Estimated Parameters for Reverse Discrimination (Attitudes) Items from the Standard MEOCS using Samejima's Graded Response Model

Item Number	<u>a</u>	<u>b₁</u>	<u>b₂</u>	<u>b₃</u>	<u>b₄</u>
RAPS 91	1.05	-1.63	-0.76	0.05	0.62
RAPS 93	1.25	-1.43	-0.46	0.29	1.02
RAPS 96	1.14	-1.92	-1.22	-0.44	0.15
RAPS 99	0.62	-3.21	-1.66	0.09	0.93
RAPS 100	1.24	-1.72	-0.96	-0.07	0.54

- RAPS 91 Some minorities get promoted just because they are minorities.
- RAPS 93 Minorities and women frequently cry "prejudice" rather than accept responsibility for personal faults.
- RAPS 96 Minorities and women get away with breaking rules that majority males are punished for.
- RAPS 99 Minorities don't take advantage of the educational opportunities that are available to them.
- RAPS 100 Many minorities act as if they are superior to majority members.

Table 15
Estimated Parameters for Reverse Discrimination (Attitudes) Items from the Standard MEOCS Revised using Samejima's Graded Response Model

Item Number	<u>a</u>	<u>b₁</u>	<u>b₂</u>	<u>b₃</u>	<u>b₄</u>
MEOCS 83	1.20	-2.04	-1.02	-0.39	0.25
MEOCS 84	1.13	-1.47	-0.30	0.44	1.19
MEOCS 85	1.25	-1.68	-0.60	0.04	0.77
MEOCS 86	0.87	-2.94	-1.52	-0.05	0.78
MEOCS 87	1.12	-2.36	-1.13	-0.18	0.45

- MEOCS-R 83 Some persons get promoted just because they are of a certain race/ethnicity.
- MEOCS-R 84 Some people cry "prejudice" rather than accept responsibility for personal

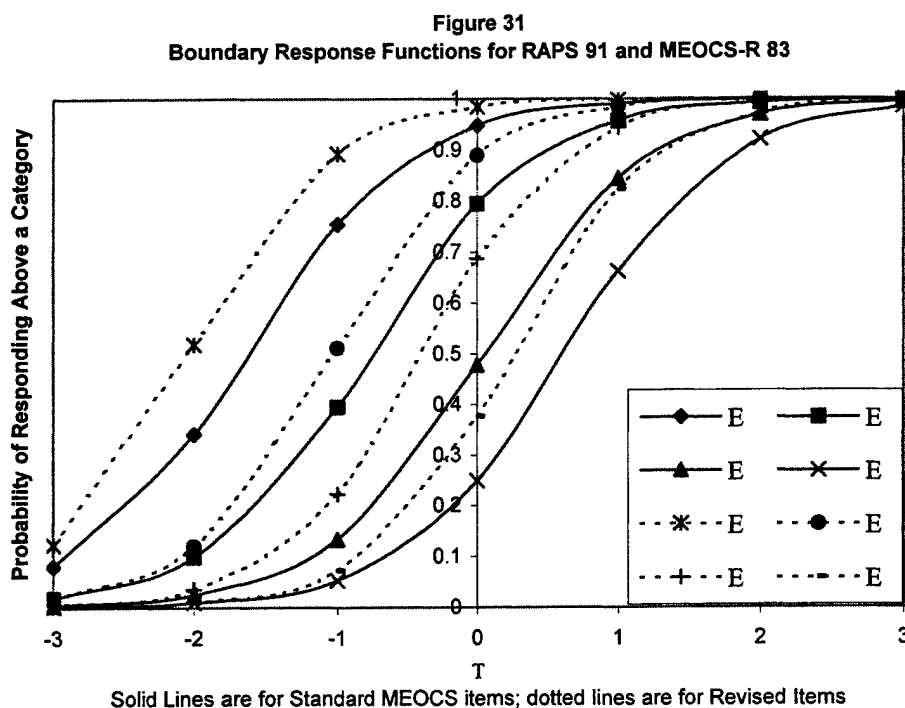
faults.

MEOCS-R 85 Some people get away with breaking rules that others are punished for.

MEOCS-R 86 People of certain racial and ethnic groups don't take advantage of the educational opportunities that are available to them.

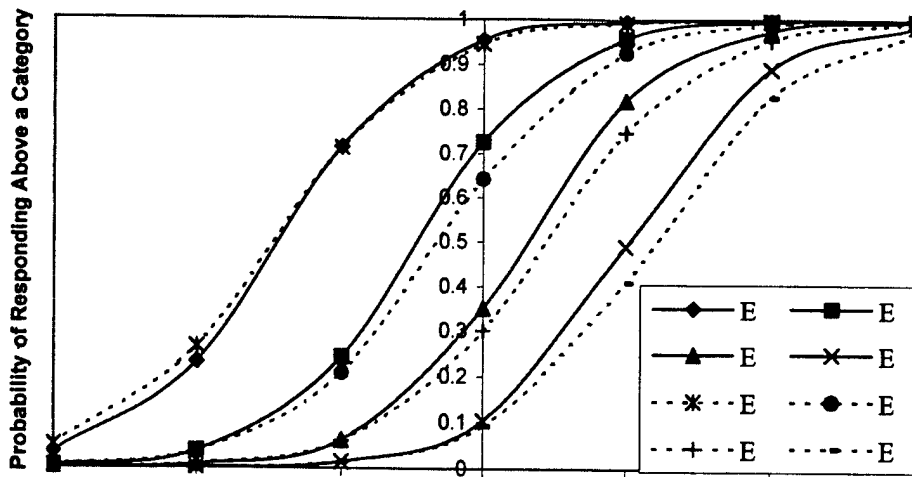
MEOCS-R 87 People of certain racial and ethnic groups act as if they are superior to others.

As can be seen in Figure 31, the BRFs for RAPS 91 and MEOCS-R 83 show DIF ($E = -.479, p < .001$). There is a uniform shift to the left with the revised item.



As can be seen in Figure 32, the BRFs for RAPS 93 and MEOCS-R 84 show a mixed pattern ($E = .178, p < .001$). For b_1 , the curves are nearly identical, for b_2 , b_3 , and b_4 there is a shift to the left.

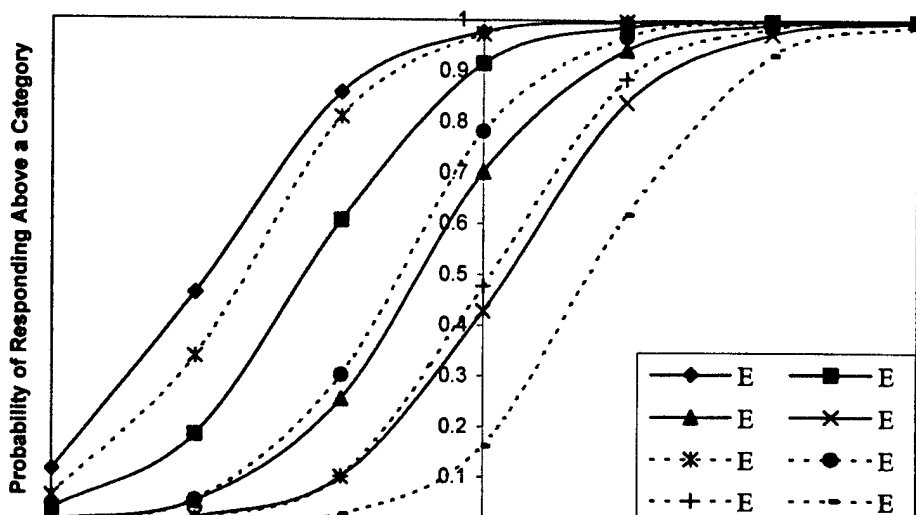
Figure 32
Category Response Functions for RAPS 93 and MEOCS-R 84



Solid Lines are for Standard MEOCS items; dotted lines are for Revised Items

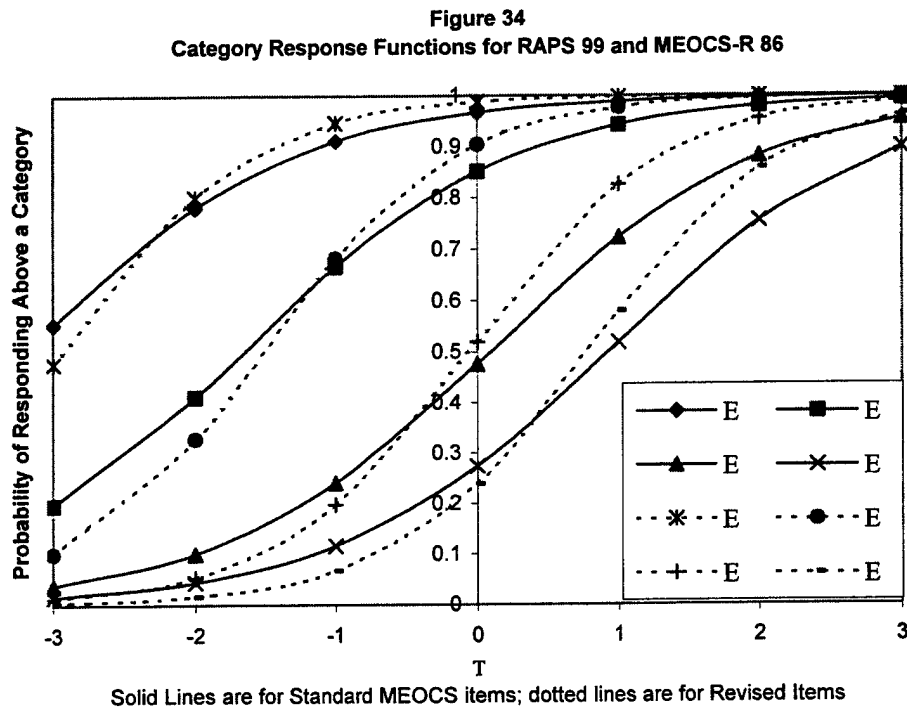
As can be seen in Figure 33, the BRFs for RAPS 96 and MEOCS-R 85 show DIF ($E = -.663, p < .001$). There is a uniform shift to the right with the revised item.

Figure 33
Boundary Response Functions for RAPS 96 and MEOCS 85



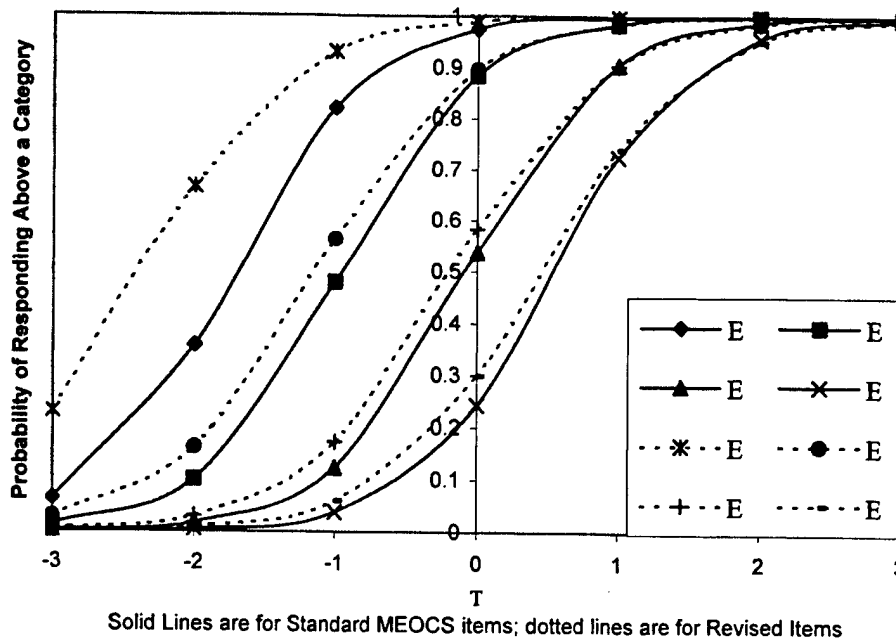
Solid Lines are for Standard MEOCS items; dotted lines are for Revised Items

As can be seen in Figure 34, the BRFs for RAPS 99 and MEOCS-R 86 show nonuniform DIF ($E = -.074$, $p = .113$). The lack of significance between these versions is probably due to the curves crossing over. For these items, the original versions are higher at low T , while the revised versions are higher at high T .



As can be seen in Figure 35, the BRFs for RAPS 100 and MEOCS-R 87 show DIF ($E = -.223$, $p < .001$). There is a uniform shift to the left with the revised item.

Figure 35
Category Response Functions for RAPS 100 and MEOCS 87



Attitudes Toward Racial/Gender Separatism

The discrimination and difficulty parameters for the original and revised versions of the Attitudes toward Racial/Gender Separatism scale are presented in Tables 16 and 17, respectively. Both versions show good discrimination and negative discrimination parameters. These discrimination parameters suggest that the items discriminate better at the lower end of the scale. The marginal reliabilities for both scales are .72 and .71, respectively. The transformation constants for Table 17 A = 1.109 and K = .219.

Table 16
Estimated Parameters for Attitudes toward Racial/Gender Separatism Items from the Standard MEOCS using Samejima's Graded Response Model

<u>Item Number</u>	<u>a</u>	<u>b₁</u>	<u>b₂</u>	<u>b₃</u>	<u>b₄</u>
RAPS 80	1.13	-2.49	-2.11	-1.22	-0.62
RAPS 82	0.89	-2.80	-1.99	-0.78	0.00
RAPS 87	1.84	-2.24	-1.74	-1.02	-0.56
RAPS 88	1.43	-2.55	-2.03	-1.11	-0.61
RAPS 92	1.11	-2.52	-1.98	-0.92	-0.37

- RAPS 80 After duty hours, people should stick together in groups made up of their race only (e.g., minorities only with minorities and majority members only with majority members).
- RAPS 82 Trying to bring about the integration of women and minorities is more

- trouble than it's worth.
- RAPS 87 Minorities and majority members would be better off if they lived and worked only with people of their own races.
- RAPS 88 I dislike the idea of having a supervisor of a race different from mine.
- RAPS 92 Power in the hands of minorities is a dangerous thing.

Table 17

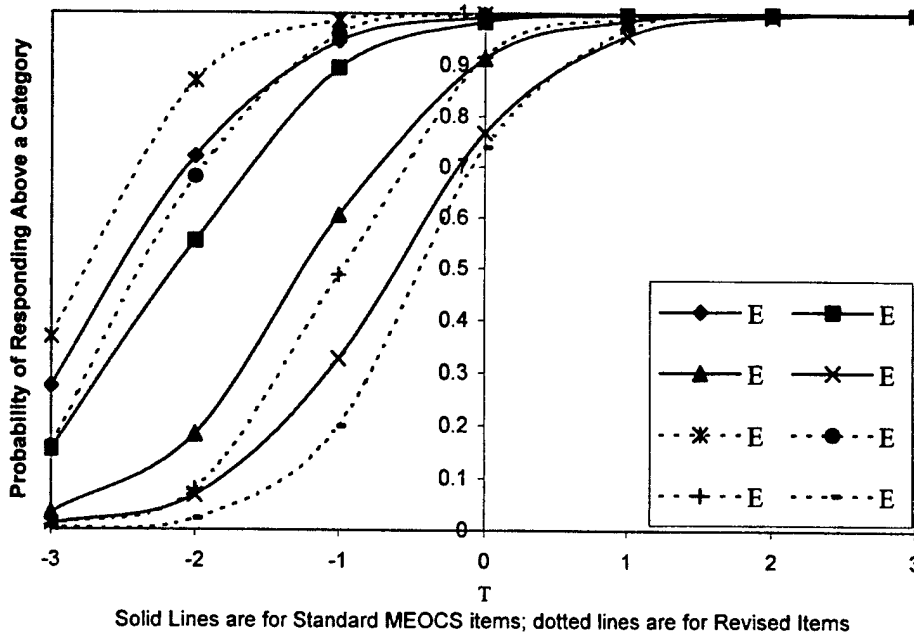
Estimated Parameters for Attitudes toward Racial/Gender Separatism Items from the Standard MEOCS Revised using Samejima's Graded Response Model

<u>Item Number</u>	<u>a</u>	<u>b₁</u>	<u>b₂</u>	<u>b₃</u>	<u>b₄</u>
MEOCS 88	1.43	-2.78	-2.31	-0.98	-0.43
MEOCS 89	1.32	-2.55	-1.77	-0.75	-0.02
MEOCS 90	2.53	-2.41	-2.03	-1.24	-0.60
MEOCS 91	1.62	-2.60	-2.15	-1.21	-0.70
MEOCS 92	0.80	-2.64	-1.70	-0.70	-0.07

- MEOCS-R 88 After duty hours, people should stick together in groups made up of their own race and ethnic group only.
- MEOCS-R 89 Trying to bring about the integration of gender, racial and ethnic groups is more trouble than it's worth.
- MEOCS-R 90 People would be better off if they lived and worked only with people of their own race or ethnicity.
- MEOCS-R 91 I dislike the idea of having a supervisor of a race or ethnic group different from mine.
- MEOCS-R 92 Power in the hands of members of some racial and ethnic groups is a dangerous thing.

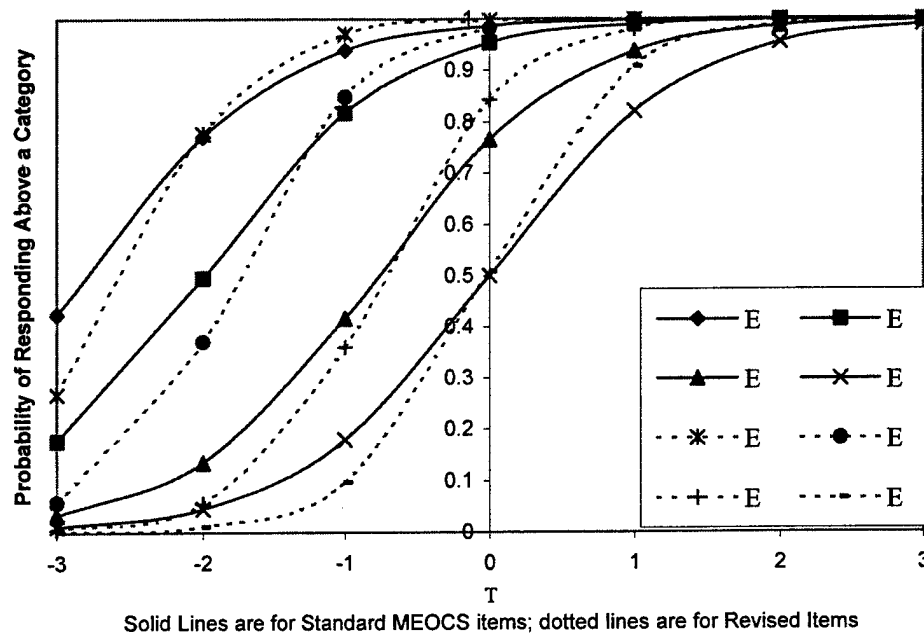
As can be seen in Figure 36, the BRFs for RAPS 80 and MEOCS-R 88 show a mixed pattern ($E = .004$, $p = .915$). The lack of significance between these versions is probably due to the different patterns for the pairs of curves. For b_1 and b_2 , there is a shift to the left. There is non-uniform DIF for b_3 and b_4 . In this case, the original version is higher at low T, while the revised version is higher at high T.

Figure 36
Category Response Functions for RAPS 80 and MEOCS-R 88



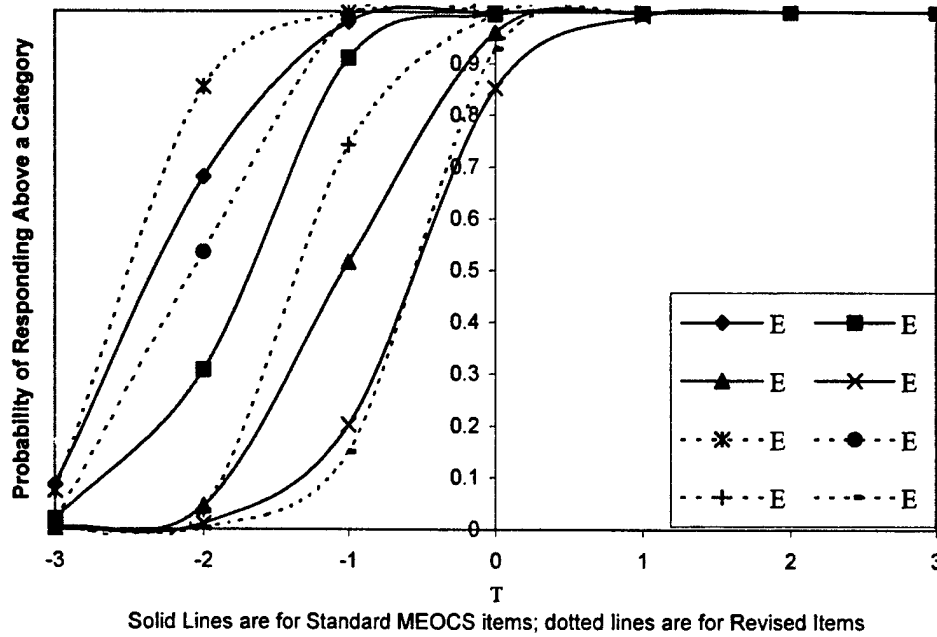
As can be seen in Figure 37, the BRFs for RAPS 82 and MEOCS-R 89 show non-uniform DIF ($E = -.062$, $p = .111$). The lack of significance between these versions is probably due to the curves crossing over. For these items, the original versions are higher at low T , while the revised versions are higher at high T .

Figure 37
Category Response Functions for RAPS 82 and MEOCS-R 89



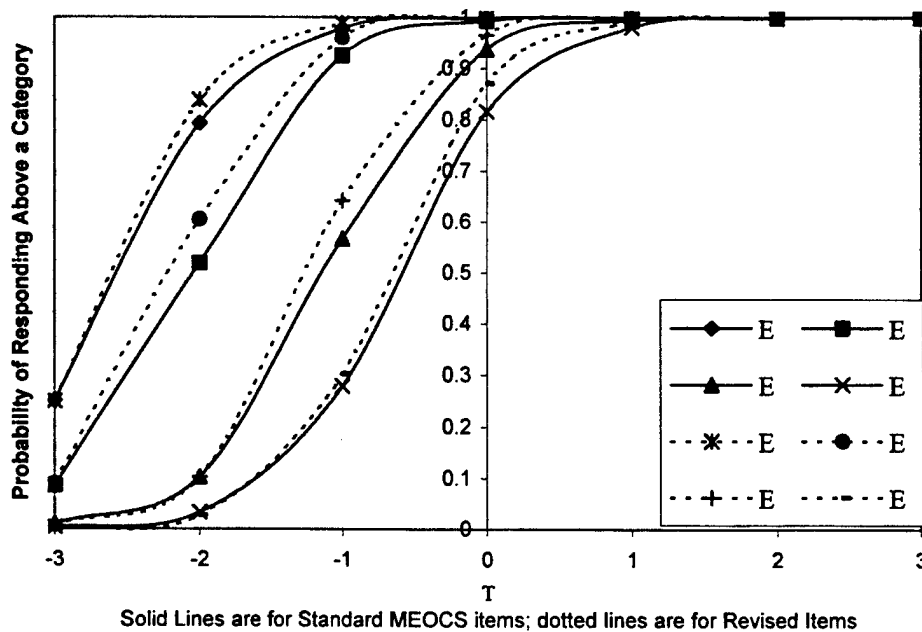
As can be seen in Figure 38, the BRFs for RAPS 87 and MEOCS-R 90 show a mixed pattern ($E = -.205$, $p < .001$). For b_1 , b_2 , and b_3 , there is a shift to the left; for b_4 , there is non-uniform DIF. In this case, the original version is higher at low T , while the revised version is higher at high T .

Figure 38
Category Response Functions for RAPS 87 and MEOCS 90

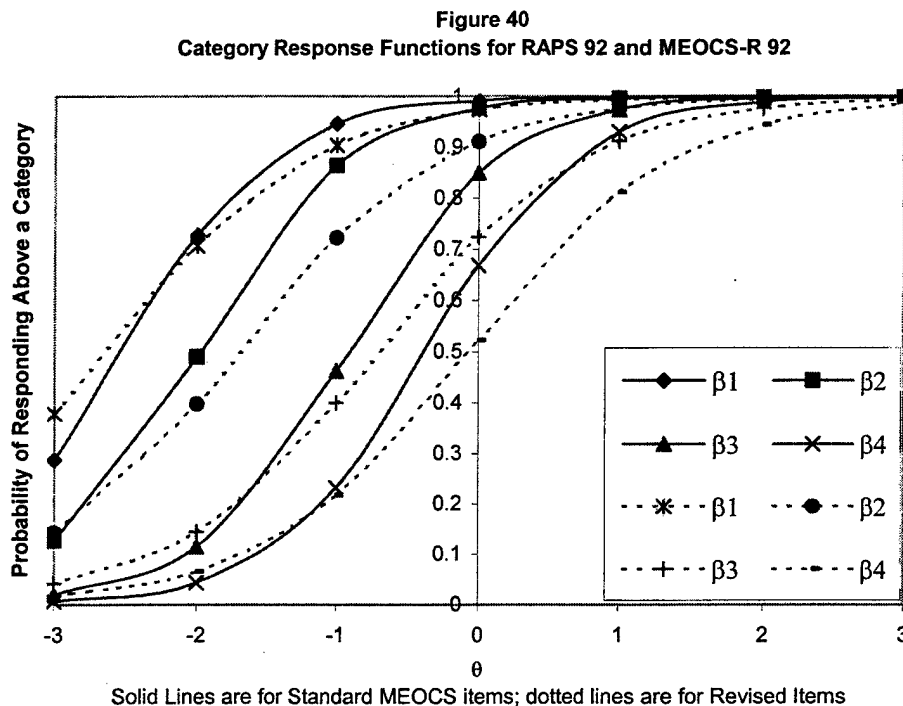


As can be seen in Figure 39, the BRFs for RAPS 88 and MEOCS-R 91 show DIF ($E = -.105, p = .001$). There is a uniform shift to the left with the revised item.

Figure 39
Category Response Functions for RAPS 88 and MEOCS-R 91



As can be seen in Figure 40, the BRFs for RAPS 92 and MEOCS-R 92 show DIF ($\beta = .352, p < .001$). For b_2 , there is a shift to the right. There is non-uniform DIF for b_1 , b_3 , and b_4 . In this case, the revised version is higher at low θ , while the original version is higher at high θ .



Discussion

Many of the analyses found that there was considerable DIF for many items. This may be viewed as a failure in the process of making the MEOCS more neutral. While lack of DIF can indicate that the neutral version of an item works as well as the original version of the item, DIF can indicate the neutral version is an improvement over the original version. Most the items on the MEOCS discriminate well between those selecting the lower end of the rating scale but do not do well in discriminating between those who select the upper end of the scale. (An exception is the Positive EO Behavior scale where the reverse is true because of the wording of the items). Thus, uniform DIF with a shift to the right, and non-uniform DIF where lower-level boundaries shift to the left and upper-level boundaries shift to the right (or a variant of this) would indicate an improvement by the neutral version of the item.

In evaluating the scales below, three DIF outcomes can be viewed positively:

- no DIF;
- uniform DIF with a shift to the right; and
- non-uniform DIF with lower-level boundaries shifting to the left and upper-

level boundaries shifting to the right (or a variant).

Two DIF outcomes can be viewed negatively:

- uniform DIF with a shift to the left; and
- non-uniform DIF with lower-level boundaries shifting to the right and upper-level boundaries shifting to the left (or a variant).

Again, it should be noted that the reverse is true when evaluating items from the Positive EO scale.

Sexual Harassment and Discrimination

Three out of the five items on this scale display negative DIF outcomes. Part of the problem may be with the nature of sexual harassment and discrimination. While it is clearly possible for men to experience sexual harassment and discrimination, respondents reading the neutral version of the item may perceive it as referring to women. Thus, the attempt to make these items more neutral may reduce their psychometric effectiveness.

Differential Command Behavior toward Minorities and Women

All five of the items on this scale display positive DIF outcomes (two with no change in DIF). When these items are written in a neutral form, their content deals with discriminatory behavior by commanding officers. Thus, majority men who feel that they have been discriminated against (so-called "reverse discrimination") can also agree with these items. As a result of the rewriting, this scale can be re-titled "Discriminatory Command Behavior."

Positive EO Behavior

All five of the items on this scale display negative DIF outcomes. The major difference in the wording is that the terms "majority" and "minority" members in the original versions have been replaced by members "of different racial/ethnic backgrounds." While abandoning terms like "majority" and "minority" may be advisable as the American society and the military become more diverse, the phrase replacing them may be too cumbersome for respondents.

Racist/Sexist Behaviors

Three out of five items on this scale display negative DIF outcomes. As in the above scale, the terms "majority" and "minority" members in the original versions have been replaced by members "of different racial/ethnic backgrounds." The replacement phrase may be cumbersome, although two of the three items where this has been done display positive DIF outcomes.

It is noteworthy that two of the items on this scale have not been rewritten, yet have negative outcomes. Perhaps a kind of overcorrection has occurred when the other three items of the scale are equated.

Reverse Discrimination (Behavior)

All five of the items on this scale display positive DIF outcomes. When these items are written in a neutral form, their content deals with discriminatory behavior. Thus, minorities and women who feel that they have been discriminated against can also agree with these items. This scale can be re-titled "Discriminatory Behavior."

Discrimination against Minorities and Women

Four of the five items on this scale display positive DIF outcomes (two with no change in DIF). It appears that neutral versions of the items work as effectively as the original versions.

Reverse Discrimination (Attitudes)

Three of the five items of this scale display negative DIF outcomes. Although it would appear that these items written in a neutral form would deal with discriminatory attitudes, the effects are otherwise. Respondents reading the neutral version of the items may perceive them as referring to women and minorities. It is noteworthy that this scale is one of the few for which the difficulty parameters are both positive and negative. Perhaps further rewriting of the neutral versions of the items is needed.

Attitudes toward Racial/Gender Separatism

Three of the five items of this scale display positive DIF outcomes. Most of the items show a slight change in revision. For example, references to "race" in the original version have been replaced by "racial and ethnic group." These revised items appear to work as effectively (if not, more so) as the original items.

General Comments

The results of this study should be viewed tentatively. The author has not had a lot of experience with the use of the SIBTEST program (Shealy & Stout, 1993). It is advisable to compare the results obtained with other DIF programs (e.g., DFIT; Raju et al., 1995). Nevertheless, the current findings are supportive of the current effort to revise items that compose the MEOCS.

The Advantages of IRT to the MEOCS

IRT can provide several advantages to the MEOCS in the areas of test construction, shortening scales, the use of DIF and DTF, and in computerized adaptive testing. With respect to test construction, researchers who are rewriting items for a revised MEOCS

should pretest these items. IRT analysis can then compare the new values for a and b with those previously established. Researchers should be aware that a lack of correspondence is not necessarily bad if a increases and b values are closer to zero. One additional advantage is that researchers can adjust the format of parts of the MEOCS to meet the requirements of the scale. Thus, if one scale is better measured in a true/false format, and another is better measured in a seven-point Likert-type scale, researchers do not have to compromise between the two.

Previous studies by Stark et al. (2002) and Truhon (2000) have shown that IRT can be used to shorten scales without a loss in psychometric qualities like reliability and validity. There has been increased concern by dropping rates of military personnel responding to the MEOCS. A shorter but just as reliable and valid version of the MEOCS may elicit a higher level of response.

DIF and DFT can help provide answers to questions often asked about the MEOCS. Whites perceive less discrimination and racism in the military than do minority groups. Men perceive less sexism in the military than do women. Senior leaders perceive fewer of these problems than do lower-rank military personnel. In each case, it is not known whether these groups see less of the same problem or if they interpret EO differently. DIF and DFT are well suited for this type of analysis.

One of the major advantages of IRT is its application to computerized adaptive testing (CAT). CAT involves presenting a test instrument by computer instead of by paper and pencil. Making use of IRT, CAT items are tailored to the characteristics of the examinee. This can result in shorter tests that provide as much psychometric information as paper-and-pencil tests.

Typically, CAT has been done with ability tests with large item banks. However, recent findings suggest that CAT could be applied to the MEOCS. Dodd, De Ayala, and Koch (1995) found the size of item bank needed is much lower when items are polytomously scored, as is done with the MEOCS. CAT has been done with several personality measures, including the Multidimensional Personality Questionnaire (Waller & Reise, 1989), the Minnesota Multiphasic Personality Inventory (Handel, Ben-Porath, & Watt, 1999), and with performance evaluation (Borman, Buck, Hanson, Motowidlo, Stark, & Drasgow, 2001).

References

- Baker, F. B. (1992). Equating tests under the graded response model. *Applied Psychological Measurement*, 16, 87-96.
- Barnes, R. D. (1996). *Toward a second generation MEOCS: Recommendations for administration format and issue coverage*. (DEOMI Research Series Pamphlet 96-12). Patrick AFB, FL: Defense Equal Opportunity Management Institute.
- Bergman, M., Palmieri, P. A., Drasgow, F., & Ormerod, A. J. (2001). Assessing racial and ethnic harassment and discrimination in diverse populations. Paper presented at the 16th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.
- Borman, W. C., Buck, D. E., Hanson, M. A., Motowidlo, S. J., Stark, S., & Drasgow, F. (2001). An examination of the comparative reliability, validity, and accuracy of performance ratings made using computerized adaptive rating scales. *Journal of Applied Psychology*, 86, 965-973.
- Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.
- Chang, H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement*, 33, 333-353.
- Collins, W. C., Raju, N. S., & Edwards, J. E. (2000). Assessing differential functioning in a satisfaction scale. *Journal of Applied Psychology*, 85, 451-461.
- Dansby, M. R., Johnson, J. L., McIntyre, R. M., & Truhon, S. A. (2001). Developing the MEOCS 2000. Panel session at the 3rd Biennial EO/EEO Research Symposium, Cocoa Beach, FL. (Published in D. McKay (ed.), *Proceedings: 3rd Biennial EO/EEO Research Symposium*. [DEOMI Research Series Pamphlet 01-1, pp. 176-198]. Patrick AFB, FL: Defense Equal Opportunity Management Institute.
- Dodd, B. G., & De Ayala, R. J., & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement*, 19, 5-22.
- Donovan, M. A., & Drasgow, F. (1999). Do men's and women's experiences of sexual harassment differ? An examination of the differential test functioning of the Sexual Experiences Questionnaire. *Military Psychology*, 11, 265-282.
- Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland and H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum Associates.

- Edwards, J. E., Elig, T. W., Edwards, D. L., & Riemer, R. A. (1997, April). *The 1995 Armed Forces Sexual Harassment Survey: Codebook for Form B (Report 95-014)*. Arlington, VA: Defense Manpower Data Center.
- Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment*, 8, 341-349.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
- Handel, R. W., Ben-Porath, Y. S., & Watt, M. (1999). Computerized adaptive assessment with the MMPI-2 in a clinical setting. *Psychological Assessment*, 11, 369-380.
- Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research*, 19, 49-78.
- Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, 9, 139-164.
- Landis, D., Dansby, M. R., & Faley, R. H. (1993). The Military Equal Opportunity Climate Survey: An example of surveying in organizations. In P. Rosenfeld, J. E. Edwards, & M. D. Thomas (eds.), *Improving organizational surveys: New directions, methods, and applications* (pp. 122-142). Newbury Park, CA: Sage.
- Li, H., & Stout, W. (1996). A new procedure for detecting crossing DIF. *Psychometrika*, 61, 647-677.
- Lord, F. M. (1952). *A theory of test scores* (Psychometric Monograph No. 7). Iowa City, IA: Psychometric Society.
- McIntyre, R. M. (1999). *A confirmatory factor analysis of the Military Equal Opportunity Climate Survey, Version 2.3*. (DEOMI Research Series Pamphlet 99-5). Patrick AFB, FL: Defense Equal Opportunity Management Institute.
- Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, 87, 517-529.
- Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement*, 19, 353-368.

- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, No. 17*.
- Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York: Springer.
- Schneider, K. T., Hitlan, R. T., & Radhakrishnan, P. (2000). An examination of the nature and correlates of ethnic harassment experiences in multiple contexts. *Journal of Applied Psychology, 85*, 3-12.
- Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*, 159-164.
- Stark, S., Chernyshenko, O. S., Chuah, D., Lee, W., & Wadlington, P. (2001). *IRT modeling lab tutorial*. Retrieved from <http://work.psych.uiuc.edu/irt>.
- Stark, S., Chernyshenko, O. S., Lancaster, A. R., Drasgow, F., & Fitzgerald, L. F. (2002). Toward standardized measurement of sexual harassment: Shortening the SEQ-DoD using item response theory. *Military Psychology, 14*, 49-72.
- Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201-210.
- Thissen, D. (1991). *MULTILOG User's Guide* (Version 6.0). Lincolnwood, IL: Scientific Software.
- Truhon, S. A. (1999). *Updating the MEOCS using cluster analysis and reliability*. (DEOMI Research Series Pamphlet 99-8). Patrick AFB, FL: Defense Equal Opportunity Management Institute.
- Truhon, S. A. (2000). *Shortening the MEOCS using item response theory*. (DEOMI Research Series Pamphlet 00-8). Patrick AFB, FL: Defense Equal Opportunity Management Institute.
- Waller, N. G., & Reise, S. P. (1989). Computerized adaptive personality assessment: An illustration with the absorption scale. *Journal of Personality and Social Psychology, 57*, 1051-1058.
- Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.